BIOCHEMICAL AND STRUCTURAL CHARACTERIZATION OF BACTERIAL RNA-GUIDED DNA TARGETING SYSTEMS

by

Hongfan Chen

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

April 2016

Abstract

The CRISPR-Cas (clustered regularly interspersed short palindromic repeats-CRISPR-associated proteins) immune systems found in many prokaryotes rely on small guide CRISPR RNAs (crRNAs) to destroy invading viruses and plasmids. This RNA-guided adaptive immune response is mediated by numerous diverse Cas proteins, several of which form complexes with crRNAs and function to silence foreign nucleic acids. Current understanding of the molecular basis of these proteins is limited. Here, we present biochemical and structural characterization of two sets of such proteins: Cascade from *Escherichia coli* and Cas9 from *Streptococcus thermophilus*.

E. coli Cascade, a large multimeric ribonucleoprotein complex, uses crRNA to base pair with complementary DNA (protospacer) at sites adjacent to a signature sequence termed the protospacer adjacent motif (PAM). The bound structure, known as an R-loop, propagates from PAM to the other end of the protospacer. A crystal structure of Cascade bound to a ssDNA target, previously determined by our laboratory, reveals a potential pocket for binding of the displaced strand in the R-loop. Here we provide experimental evidence that this pocket serves as a docking site for the displaced strand, and this binding facilitates DNA strand separation during R-loop formation. Structure-guided mutagenesis of the basic residues in the pocket confirms their importance for double strand DNA binding. Single-molecule experiments reveal that these mutations kinetically hinder R-loop formation. We further show that Cascade exerts a strong conformational "lock" upon completion of an R-loop, and this locked conformation is sufficient for recruiting the trans-acting Cas3 helicase/nuclease for target destruction.

Cas9 from *S. thermophilus* LMG18311 is 1122 amino acid protein harboring a HNH nuclease domain and a RuvC-like nuclease domain. We demonstrate that LMG18311 Cas9 utilizes a crRNA in conjunction with a trans-acting crRNA (tracrRNA) to cleave double strand DNA *in vivo* and *in vitro*. The cleavage is dependent on the presence of PAM as well as the position of the PAM. We further show that the HNH and RuvC-like nuclease domains of Cas9 select the location of their cleavage sites via different mechanisms. The HNH domain catalyzes cleavage of the target strand at a fixed position, whereas the RuvC-like domain catalyzes cleavage of the non-target strand using a ruler mechanism.

Advisor:

Dr. Scott Bailey

Thesis Readers:

Dr. Paul Miller

Dr. Sean Prigge

Dr. Jungsan Sohn

Alternative Readers:

Dr. Roger McMacken

Dr. Srinivasan Chandrasegaran

Acknowledgements

My graduate school has been an experience of scientific exploration as well as personal growth. It has taught me the importance of persistence, self-belief, and appreciation. Looking back, all the doubles and struggle fade away, and I truly believe that I could not have made it without people who helped me along the path.

First and foremost, I would like to express my gratitude to my advisor, Dr. Scott Bailey, for his guidance and mentorship in the past five years. He invested tremendous amount of time and effort in my projects. He is always available for discussion and questions, which greatly accelerated my research progress. His rigorous and down-to-earth approach towards science helped shape my way of thinking and doing science. Apart from that, I am especially grateful that he gave me freedom to explore my other career interests.

I would like to thank my wonderful co-workers and collaborators who have made my work possible. In particular, my thanks go to current Bailey lab members John Mallon, Jasvir Kaila, past members Sabin Mulepati, John Choi, and our collaborators Dr. Ralf Seidel and Dr. Christophe Rouillon (Universität Leipzig, Germany), who worked on different aspects of the work presented here. I thank all other Bailey lab members whom I am fortunate to cross my path with for making the lab a very enjoyable workplace. I also want to thank for all other BMB laboratories for their generosity with reagents and equipment.

I would like to thank my thesis committee, Drs. Sean Prigge, Paul Miller, and Jungsan Sohn for thoughtful advices during thesis meetings and careful reading of my thesis.

I am grateful that I have made a few good friends during graduate school who believed in me and helped me push through. I want to thank Jingchuan Luo and Allen Cheng for always being there for me and for all the cheering and support they gave me whenever I need it. Daisy Colón-López has being a good friend throughout my graduate school and helped me with my life in every aspect. Lastly, I cannot thank my boyfriend Gaoran Yu enough for his constant encouragement and great sense of humor that brightens my life.

Last but not least, I am greatly indebted to my parents for their unwavering love, support and understanding. When I decided to come to US for graduate school, I did not realize how much I would be giving up. But they supported me along the way no matter what happened. Without them, none of these would be possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ATP	adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	base pair
BSA	bovine serum albumin
Cas	CRISPR-associated
Cascade	CRISPR-associated complex for antiviral defense
CFU	colony forming units
CRISPR	clustered regularly interspersed short palindromic repeats
crRNA	CRISPR RNA
CTD	carboxyl-terminal domain
CTP	cytidine triphosphate
DNA	deoxyribonucleic acid
ds	double-strand
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
EM	electron microscopy
GTP	guanosine triphosphate
HIV-1	human immunodeficiency virus type 1
IMAC	immobilized metal affinity chromatography
IPTG	isopropyl β-D-1-thiogalactopyranoside
K_d	dissociation equilibrium constant
kD	kilodalton

LB	Luria-Bertani
MBP	maltose-binding protein
MCS	multiple cloning site
nt	nucleotide
NUC	Cas9 nuclease lobe
PAGE	polyacrylamide gel electrophoresis
PAM	protospacer adjacent motif
PCR	polymerase chain reaction
PEG	polyethylene glycol
REC	Cas9 recognition lobe
RNA	ribonucleic acid
SDS	sodium dodecyl sulfate
SeMet	selenomethionine
sgRNA	single guide RNA
ss	single-strand
SSC	spermatogonial stem cells
TCEP	tris(2-carboxyethyl)phosphine
TEV	tobacco etch virus
tracrRNA	*trans*-activating crRNA
UTP	uridine triphosphate
WT	wild-type
β-ME	2-mercaptoethanol

Chapter 1

Overview of the CRISPR-Cas Systems

Introduction to CRISPR-Cas systems

Viruses are the most abundant life form on earth, outnumbering their hosts by ten fold (1). Bacteria and archaea are under constant threat from prokaryotic viruses (bacteriophages or phages). As a consequence, multiple defense mechanisms have evolved in prokaryotes to combat phage infections. Among them, mechanisms to inhibit phage adsorption, restriction-modification, and abortive infection systems are some better characterized antiviral strategies which provide innate immunity (2). On the other hand, a recently identified CRISPR-Cas (clustered regularly interspersed short palindromic repeats-CRISPR-associated proteins) system represents a distinct defense mechanism that provides adaptive immunity against phage infection as well as plasmid conjugation (3-7). The CRISPR-Cas system is widespread; it is present in most archaea and about half of the bacteria genomes (8).

A CRISPR-Cas locus comprises two components: a CRISPR array and a cas operon. The CRISPR array consists of a stretch of identical short direct repeats (25-40 base pairs (bp)) separated by unique invader-derived "spacer" sequences of similar length (Mojica et al. 2005; Bolotin 2005; Pourcel et al. 2005). A set of *cas* genes is usually located adjacent to the CRISPR loci and encodes Cas proteins that play essential roles in the CRISPR-Cas activity. The immunity mediated by CRISPR-Cas is carried out in three stages (Fig 1.1): acquisition of new spacers, generation of small guide CRISPR RNA (crRNA), and target interference. In the first stage, prokaryotes adapt to invasive mobile elements, such as phages and plasmids, by taking up short DNA fragments (protospacers) and integrating into the CRISPR locus as new spacer sequences (5,6). These spacer sequences record past exposures and will be utilized against future invasions from the

same genetic elements. Secondly, the CRISPR locus is transcribed into a primary transcript and processed to generate mature crRNAs that contain spacer-derived sequences. Finally, the crRNA, assembled into an effector complex with Cas protein/s, directs interference with target nucleic acids in a sequence-specific manner. This usually results in target cleavage by the Cas nucleases. As such, CRISPR-Cas is able to mount a rapid and robust counterattack against invading genetic elements in prokaryotes.

The CRISPR-Cas systems are rapidly evolving. Apart from the universally conserved Cas1 and Cas2 that are responsible for spacer acquisition, other Cas proteins and their functions are vastly diverse (8,9). Based on the configuration of the effector complexes, current classification broadly divides CRISPR-Cas systems into two classes: class 1 utilizes multiple proteins complexed with crRNA to mediate target recognition and degradation, whereas class 2 employs a single large protein in conjunction with crRNA to fulfill the function (8) (Fig 1.2). Class 1 systems consist of type I, type III and a putative new type IV, and Class 2 systems consist of type II and a putative new type V (Fig 1.2) (8). Type I, II and III were previously defined according to the presence of their signature proteins: Cas3 for type I, Cas9 for type II, and Cas10 for type III, respectively (10) (Fig 1.2); each of these systems can be further divided into subgroups. A detailed review of type I and type II systems, with an emphasis on target interference, is presented below.

Type I systems

Type I systems are the most prevalent CRISPR-Cas systems among both bacteria and archaea (8). Typical type I mediated immunity requires the multi-subunit effector complex termed Cascade (CRISPR-associated complex for antiviral defense) and the translocating nuclease Cas3. A great body of knowledge of type I systems comes from structural and biochemical studies of *Escherichia coli* type I-E (Fig 1.3). Firstly, Cas1 and Cas2 comprise a hexameric intergrase complex that uses a cut-ant-paste mechanism to incorporate new spacers into the CRISPR array (11,12). In some cases, if a variant of a pre-existing spacer is encountered, the interference components Cascade and Cas3 assist Cas1 and Cas2 in spacer acquisition (13,14). This process, known as "priming", is more complex yet more effective (15). Next, the Cas6 endonuclease, a subunit of the Cascade complex, processes the primary crRNA transcript to produce a 61-nucleotide (nt) mature crRNA (16-18). The mature crRNA contains a 32-nt spacer sequence flanked by an 8-nt 5' end handle and 21-nt 3' end hairpin structure (19). Lastly, eleven protein subunits of five Cas proteins ($Cse1_1$, $Cse2_2$, $Cas5_1$, $Cas6_1$, $Cas7_6$) assemble with the crRNA to form a Cascade surveillance complex, which recognizes cognate foreign DNA and recruits Cas3 for target degradation (16,19).

Cascade

The overall structure of *E. coli* Cascade was first revealed by electron microscopy (EM) (19-21). Three subsequent atomic-resolution crystal structures advanced the understanding of this complex (22-24). Overall, the *E. coli* Cascade resembles a seahorse-like architecture (Fig 1.4A and C) in which the 3' hairpin and 5' handle of

crRNA are capped by Cas6 (head) and Cas5 (tail), respectively, and the spacer sequence is spanned along the helical arrangement formed by six Cas7 subunits (spine). The tail is further extended by the large subunit Cse1. The two small subunits Cse2 form a homodimer, positioned at the inner surface of the Cas7-crRNA spine (belly). Interestingly, in the pre-target bound form (22,24) (Fig 1.4A), the spacer sequence of crRNA is displayed as six 5-nt segments with every sixth base flipped out. The bases from the 5-nt segments are extended outwards, poised for interactions with the target DNA strand, whereas the every sixth base is kinked in an angle not suitable for base pairing. These kinks are introduced by the protruding long β-hairpins from Cas5 and Cas7.2 to 7.6. This observation is consistent with other data showing that target DNA mismatches are readily tolerated at every sixth position (25).

The cryo-EM reconstruction of Cascade before and after binding to a single-strand (ss) RNA target mimic revealed a concerted conformational rearrangement where Cas6, Cse2 dimer and Cse1 shift and rotate along the Cas7-crRNA spine (20). A 3-Å crystal structure of Cascade bound to a ssDNA target (Fig 1.4B) confirmed this observation (23). In addition, this crystal structure further showed that the crRNA:ssDNA heteroduplex adopts a highly distorted A-form architecture that mimics a ribbon. This unusual configuration is due to the interruptions from the long β-hairpins of Cas5 and Cas7 that prevent base paring at every sixth position. The underwound duplex ensures continuous base paring at each 5-nt segments. Such distorted conformation and small increments likely serve as a step-wise proofreading mechanism, because nonspecific targets that base pair incorrectly with the crRNA can not overcome the energetic cost of

the disfavored conformational change, especially at early steps, and thus will be rejected accordingly.

Cas3

Being the signature protein of type I systems, Cas3 typically consists of a HD nuclease and a Superfamily 2 helicase. Biochemical studies showed that Cas3 of type I-E, upon recruitment by Cascade-DNA binding, nicks the displaced strand and proceeds to unwind duplex DNA in a 3' to 5' direction using an "inchworm" mechanism (26,27). The nuclease activity requires Mg^{2+} or transition metal ions such as Mn^{2+} and Ni^{2+}, and the helicase activity is dependent on ATP hydrolysis (26-28). Two crystal structures of full-length type I-E Cas3 have been solved to date, one from *Thermobaculum terrenum* with and without ATP analog bound (29), and one from *Thermobifida fusca* in complex with a ssDNA (30). Both structures showed two metal ions coordinated by conserved residues in the HD nuclease domain. ATP binding at the interface of Superfamily 2 helicase RecA-like domains induces conformational changes in a motif V, suggesting it may be involved in coupling ATPase activity with ssDNA binding and translocation (29). The ssDNA-bound Cas3 structure implies that ssDNA substrate is likely fed into the catalytic site of the HD nuclease by the Superfamily 2 helicase (30). In addition, a low-resolution EM structure suggested that Cas3 is recruited by Cse1 at the PAM proximal end of the displaced strand (21). However, the detailed mechanism of Cas3 recruitment is unknown.

PAM recognition

Another important prerequisite for target recognition is the presence of a short consensus sequence adjacent to the target sequence, termed protospacer adjacent motif (PAM) (31). The acquisition machinery selectively incorporates spacers near a PAM sequence (14,31). Thus, the presence of PAM signals a non-self DNA, as the genomic CRISPR locus lacks a PAM (32). *E. coli* PAM has a consensus sequence 5'-CWT-3' (where W is an adenosine or thymidine) (31). However, recent studies found that up to 5 different PAM sequences can be recognized by Cascade for target interference, and about 22 are functional for priming, making the PAM interaction promiscuous (25). Previous work has implied that Cse1 is important for PAM recognition and Cas3 recruitment (21,33). Mutations in a loop from Cse1 impaired PAM binding (33). Additionally, the base pairing potential of PAM region affects target degradation by Cas3 (21). However, a complete understanding of PAM interaction and its role in Cas3 activation awaits detailed structural and biochemical analysis.

Type II systems

Type II systems constitute a minority of all CRISPR-Cas systems, and they appear to be present exclusively in bacteria (8). Notably, type II systems differ mechanistically from type I and type III, and have by far the minimal components. Type II *cas* operons typically comprise three to four *cas* genes, including *cas1*, *cas2*, the signature gene *cas9*, and in some cases *csn2* or *cas4* (Fig 1.2 and 1.5) (8). In addition to the *cas* operon and CRISPR array, type II loci include an atypical gene that encodes for *trans*-activating crRNA (tracrRNA), which is necessary for co-processing of crRNA and proper anchoring of crRNA in Cas9 (Fig 1.5).

Type II systems have evolved a distinct enzymatic pathway. During spacer acquisition, besides the functionally conserved Cas1 and Cas2, Cas9 and Csn2 have been recently shown to be also required in type II-A (34,35). Specifically, Cas9 appears to influence spacer selection by specifying PAM sequences (34). The guide crRNA is first cleaved by endogenous RNase III in the partially complementary repeat region of crRNA and tracrRNA, and subsequently trimmed by an unknown nuclease (36) (Fig 1.5). The resulting mature crRNA contains ~20 nt spacer sequence (36). Cas9, the only protein component required for target interference, forms a ternary complex with crRNA:tracrRNA duplex, which identifies complementary target DNA in a PAM dependent manner and introduces a blunt-end double-strand (ds) DNA break (7,37,38). The tracrRNA:crRNA duplex structure is required to activate the nuclease activities of Cas9 (36,37).

The two RNAs can be artificially fused into one single guide RNA (sgRNA), which enabled easy implementation of CRISPR-Cas9 as a genome engineering tool (37).

By designing the sgRNA, Cas9 nucleases can be programmed to target virtually any sequence of interest in the genome. The dsDNA breaks generated by Cas9 can be repaired by homologous recombination or non-homologous end joining, which can lead to different types of mutations (39-41). In addition, nuclease dead versions of Cas9, which bind to but do not cleave the target, have been employed to effectively regulate gene transcription (42). However, current research endeavors have been largely based on the *Streptococcus pyogenes* Cas9, and it has reached several limitations. While the general mechanism of Cas9-mediated DNA targeting has been unraveled, to develop an ideal Cas9-based tool for genome engineering and potentially gene therapy requires further studies of this system.

Cas9

Despite high sequence diversity, the type II signature protein Cas9 typically contains a HNH nuclease domain, a RuvC-like nuclease domain, and a characteristic Arginine-rich cluster (8) (Fig 1.6A). The HNH and RuvC-like nuclease domains have been shown to cleave DNA complementary and non-complementary strands, respectively (37,38). Crystal structures of Cas9 from *S. pyogenes* (type II-A) (Fig 1.6A and B) and a smaller ortholog from *Actinomyces naeslundi* (type II-C) both revealed a bi-lobed architecture with HNH, RuvC, and carboxyl-terminal domain (CTD) comprising the nuclease (NUC) lobe and an α-helical domain forming the other lobe, later referred to as the recognition (REC) lobe (43,44). A recent Cas9 structure of another smaller ortholog from *Staphylococcus aureus* (type II-A) revealed a similar bi-lobed conformation, suggesting the structural conservation for all Cas9 enzymes (45). The HNH domain likely

uses a one-metal ion catalytic mechanism, whereas the RuvC domain uses a two-metal ion mechanism, based on structural similarities to known nucleases (37,44,45). Comparison with sgRNA or sgRNA:DNA (Fig 1.6C and D) bound structures suggested that Cas9 adopts an auto-inhibited conformation in its apo form and undergoes conformational changes upon guide RNA binding, and further rearranges upon DNA target binding (37,44,46) (Fig 1.6B-D).

A structure of *S. pyogenes* Cas9 in complex with sgRNA (Fig 1.6C) indicated that guide RNA binding triggers Cas9 to reach a target-recognition competent conformation (46). Specifically, guide RNA binding induces structural rearrangement of the NUC and REC lobe to form a central groove to accommodate the repeat:antirepeat structure and the spacer region of sgRNA. Furthermore, the first 10-nt of the sgRNA spacer region is pre-organized in an A-form conformation, reminiscent of the RNA positioning in Argonaute proteins, suggesting they might use a similar mechanism to engage a DNA target. Structures of Cas9-sgRNA bound to DNA targets further provided insights into target recognition by Cas9 (Fig 1.6D) (44,47). Overall, the structures revealed additional conformational rearrangement of Cas9 from its guide RNA bound state, especially in the NUC lobe. The sgRNA;ssDNA heteroduplex adopts an A-form helical structure, positioned at the central groove between the NUC and REC lobe, consistent with sgRNA bound structure. The conserved Arginine-rich cluster on a helix bridging the two lobes is critical for sgRNA:DNA recognition. Comparison between these structures suggested a high degree of structural flexibility in Cas9.

PAM recognition

Similar to Cascade, PAM recognition is also a critical aspect for Cas9-mediated DNA targeting. Cas9 searches DNA for PAM sequences, and the identification of PAM triggers base paring between crRNA and target DNA in a unidirectional manner (48). *S. pyogenes* Cas9 recognizes a 5'-NGG-3' PAM. Comparison between the apo Cas9 and the sgRNA-Cas9 structures showed that guide RNA binding triggers the PAM recognition region to become ordered. A structure of *S. pyogenes* Cas9 bound to a sgRNA and a dsDNA target containing a 5'-TGG-3' PAM provided a close-up view of PAM recognition (Fig 1.6D) (47). Consistent with biochemical data, PAM is read out from the non-target strand at the GG position (37). The guanine dinucleotide is recognized in the major groove by conserved Arg residues from the CTD domain. Additional analysis revealed that PAM binding results in dsDNA melting immediately upstream of PAM, in agreement with single-molecule data (48). Structural comparison with *S. aureus* Cas9 in complex with sgRNA and dsDNA containing its consensus PAM (5'-NNGRRT-3') suggested that the specific residues in the PAM recognition region determine the distinct PAM specificities of Cas9 orthologs (45). These studies furthered the understanding of the structural basis of PAM recognition, and will help with the design of versatile Cas9 enzymes with tailored PAM sequences.

CRISPR-Cas and Public Health

Recent breakthroughs in discovering the molecular underpinning of CRISPR-Cas systems have paved the way for several applications. In particular, Cas9, the endonuclease that mediates target interference in type II systems, has attracted great attention as a genome engineering tool (39-41). Due to the simple requirements of the targeting machinery and the programmable targeting capability, Cas9-based techniques have been exploited for a number of emerging biological and medical applications (49-51). However, the use of the versatile CRISPR-Cas systems is not limited to genome engineering. Here, a few public health related perspectives of CRISPR-Cas applications are highlighted.

Pathogen typing and subtyping

Rapid and accurate identification of the source of infection is key to a timely response to a microbial outbreak. Differentiation of isolates at a sub-species level, or subtyping, presents a challenge to molecular epidemiologists (52). CRISPR provides the possibility for a precise and quick molecular typing method of bacterial pathogens. This is due to the fact that the spacer sequences in bacterial CRISPR loci are incorporated in a polarized manner: the new spacers are added to the 5' end, or the "leader sequence", of the CRISPR array (5,53). The older spacers, located distal to the leader sequence, may be shared among common ancestors. Thus, the spacer sequences can be used to trace the evolution of a particular strain. Furthermore, microvariations between strains, such as duplication or loss of internal spacers, or single nucleotide polymorphisms in spacers, provide additional information for the discrimination below the serotype level (54).

CRISPR-based typing methods have been applied for strain discrimination in *Salmonella,* a genus of bacteria that is commonly associated with foodborne illness outbreaks (54-56). Analysis of the distribution of over 3,800 unique spacers in *Salmonella* isolates showed that spacer content was strongly correlated with both serotype and the multilocus sequence typing type (54). A CRISPR-based high-throughput subtyping method has been developed for the most prevalent *Salmonella* serotype, Typhimurium (54). This method is fast, convenient, and highly specific when tested in several different *Salmonella* Typhi strains from diverse genetic and geographical origins. These studies demonstrated that CRISPR sequence analysis could be utilized as a powerful approach for typing and subtyping of human pathogens, and is likely to be of benefit to the public health laboratories.

Antibiotic Resistance

Bacterial antibiotic resistance has emerged as a serious public health threat. The prevalence of antibiotic resistance is largely due to the spreading of antibiotic genes between bacteria, a process known as horizontal gene transfer (54). This can occur through phage transduction, plasmid conjugation, and DNA transformation (57). CRISPR-Cas loci are estimated to exist in about half of human bacterial pathogens, such as *Staphylococci epidermidis*, *Streptococcus agalactiae*, *Escherichia coli* and *Helicobacter pylori,* etc. (57). However, the robustness of CRISPR-Cas immunity and the emergence of antibiotic resistance appear to be negatively correlated. CRISPR-Cas has been shown experimentally to limit horizontal gene transfer events and hence the acquisition of virulence or antibiotic resistance genes (58). In addition, a sequence

analysis of Enterococci revealed an inverse correlation between the presence of CRISPR-Cas loci and the accumulation of antibiotic resistance genes (57,59). The loss of CRISPR-Cas loci has also been reported in several other bacterial pathogens (60). These studies suggest that antibiotic use may have driven the selection towards a compromised immune system in bacteria in order to gain beneficial traits for their survival, such as antibiotic resistance genes.

Recently, with the development of CRISPR-Cas9 technology, the concept of CRISPR-based next-generation antimicrobials has been introduced (61). The basic principle is to program Cas9-crRNA to selectively target bacteria harboring specific resistance genes. Two recent studies demonstrated the feasibility of this strategy (62). Bikard *et al*. developed a phagemid system that delivers Cas9 and its guide RNA sequences to selectively kill antibiotic resistant strains of *Staphylococcus aureus*. A phagemid is a plasmid that is capable of being packaged into phage particles. Incubating the phage particles with the *S. aureus* cells resulted in a survival rate of 10^{-4}. This treatment could also immunize the antibiotic sensitive population against plasmid transfer, preventing the spread of antibiotic resistance genes. A similar strategy used by Citorik *et al*. to target antibiotic-resistant strains of pathogenic *E. coli* achieved similar results. These studies showed that it is promising to selectively kill or re-sensitize antibiotic-resistant bacteria to antibiotics by targeting specific resistance genes using the CRISPR-Cas9 technology.

Antiviral therapy

Chronic viral infections affect billions of people worldwide, and there is a dire need for a curative antiviral therapy. Current antiviral therapies are effective at targeting replicating virus, but they do not eradicate latently integrated or non-replicating proviral DNA (63,64). The advance with CRISPR-Cas9 technology has provided the possibility to directly target the integrated viral DNA, with the hope of eradicating the entire population of viral DNA genomes (65). The first study that explored the antiviral therapy potentials of CRISPR-Cas9 technology was carried out by Ebina *et al*. in human immunodeficiency virus type 1 (HIV-1) (66). Cas9 and its guide RNA were designed to target long terminal repeat sequence of HIV-1, which led to significant inhibition of virus expression and removal of the integrated proviral DNA from the cellular genome (67). Using a similar approach, Hu *et al*. showed that, in addition to the excision of latent proviruses, CRISPR-Cas can also immunize cells against HIV-1 infection (67). This work has been extended by others to remove or inactivate DNA genomes of hepatitis B virus, human papilloma virus, and herpes simplex virus with varying degrees of success (68) Although a targeted curative antiviral therapy is still in its infancy, CRISPR-Cas opens up potential new approaches to eradicate viral infections.

Cancer therapy

Cancer has significant impact on public health in the United States. Given that cancer is a genetic disorder, CRISPR-Cas9 can be harnessed to correct the oncogenic mutations or modulate epigenetic states. Early attempts using CRISPR-Cas9 to correct disease alleles have been reported by several groups in animal models of human disease

(66). For example, Xue *et al*. delivered a plasmid encoding Cas9 and sgRNA via hydrodynamic injection to directly target the tumor suppressor genes *Pten* and *p53* in mouse liver, which resulted in (~20%) hepatocyte modification (69). This study demonstrated the feasibility of generating somatic cancer mutations in adult animals using CRISPR-Cas9, which could lead to fast development of animal cancer models. In a recent study, Wu *et al*. transfected spermatogonial stem cells (SSC) with a plasmid expressing Cas9 and sgRNA, targeting a disease-causing mutation in *Crygc* that pre-existed in SSCs (70). After spermatogenesis in male mice, single SSCs that carry the desired gene correction without additional unwanted genomic changes were selected and injected into mature oocytes in female mice. This approach yielded offspring with the corrected phenotype at an efficiency of 100%, and provided proof-of-concept data of curing a genetic disease via CRISPR-Cas9 mediated gene correction in SSCs. Although these studies were carried out in animal models, they showed compelling early evidence that CRISPR-Cas9 may be deployed as a gene therapy for genetic diseases in the future.

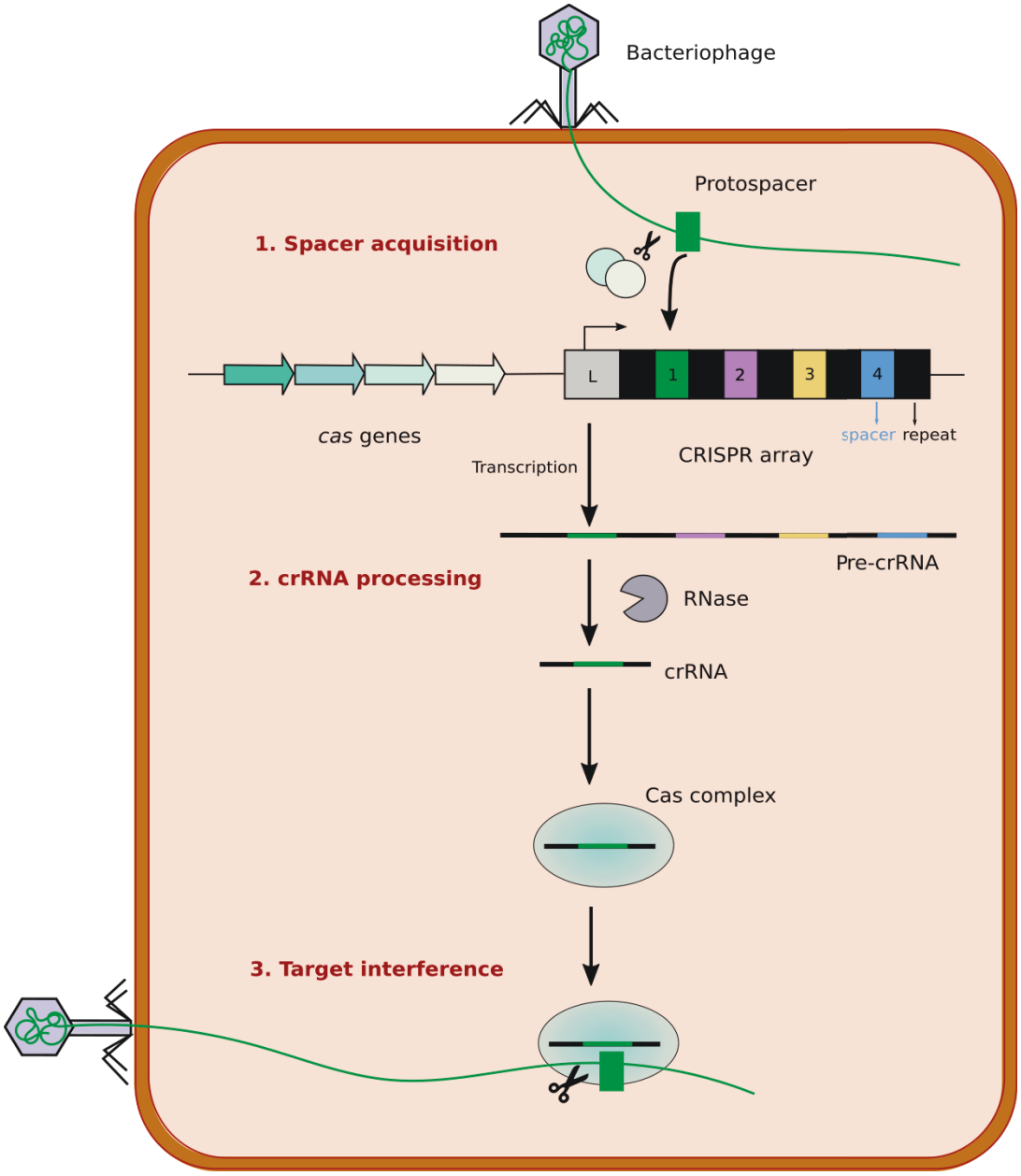

Fig 1. 1 Three stages of CRISPR-Cas immunity. A typical CRISPR locus consists of a CRISPR array and adjacent CRISPR-associated (*cas*) genes (colored arrows). The CRISPR array is composed of a leader sequence (grey box) followed by a series of identical repeats (black boxes) and unique spacers (colored boxes). In the spacer acquisition stage, short segments, called protospacers (green box), are cut (scissors) from the invading nucleic acid and incorporated into the leader proximal end (No. 1) of the CRISPR array. In the crRNA processing stage, CRISPR array is transcribed in to a long primary transcript (pre-crRNA), which undergoes processing and produces mature crRNA. During target interference, crRNA assembles with Cas proteins to form an effector complex, which identifies and cleaves (scissors) the invading nucleic acid.

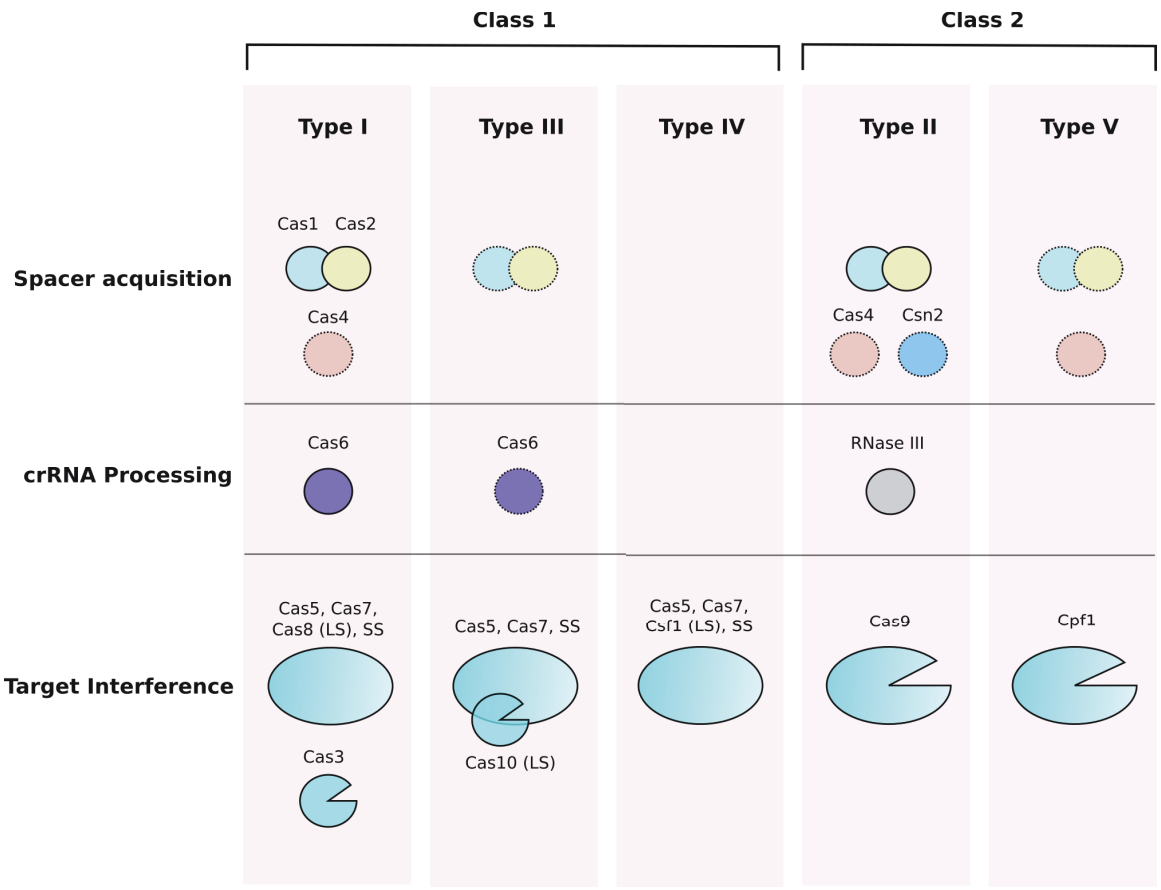

Fig 1. 2 Classification of CRISPR-Cas systems. Cas proteins are grouped into the three functional stages. Functions of Type IV and type V system components are based on homology to similar components of other systems, and have not yet been proven experimentally. Circles with dashed lines indicate that the components are not found in all subtypes within the given type. LS and SS stand for large subunit and small subunit in the effector complex. The SS Cas protein names are not shown because the nomenclature differs between subtypes. RNase III is an endogenous protein not encoded by a cas operon.

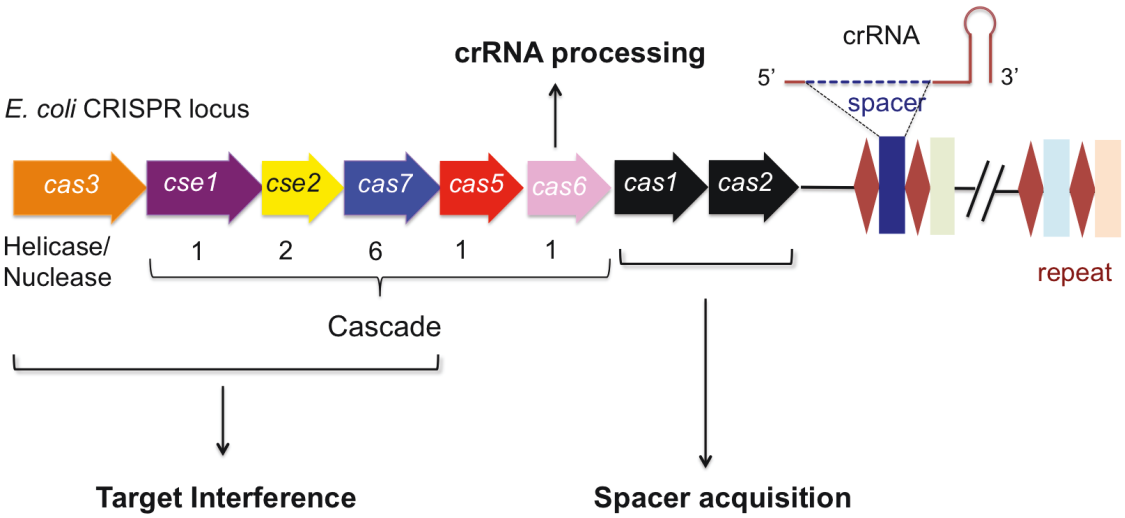

Fig 1. 3 The *E. coli* CRISPR locus and the functions of the Cas proteins. Cse1 (Cas8), Cse2, Cas7, Cas5, Cas6, with a stoichiometry of 1:2:6:1:1, together with a crRNA, constitute the CRISPR-associated complex for antiviral defense (Cascade). The colored rectangles represent different spacers, and the orange diamonds represent identical repeats. The *E. coli* crRNA contains a 32-nt spacer sequence (blue dashed line) flanked by an 8-nt 5' end handle (orange line) and 21-nt 3' end hairpin structure (orange line).

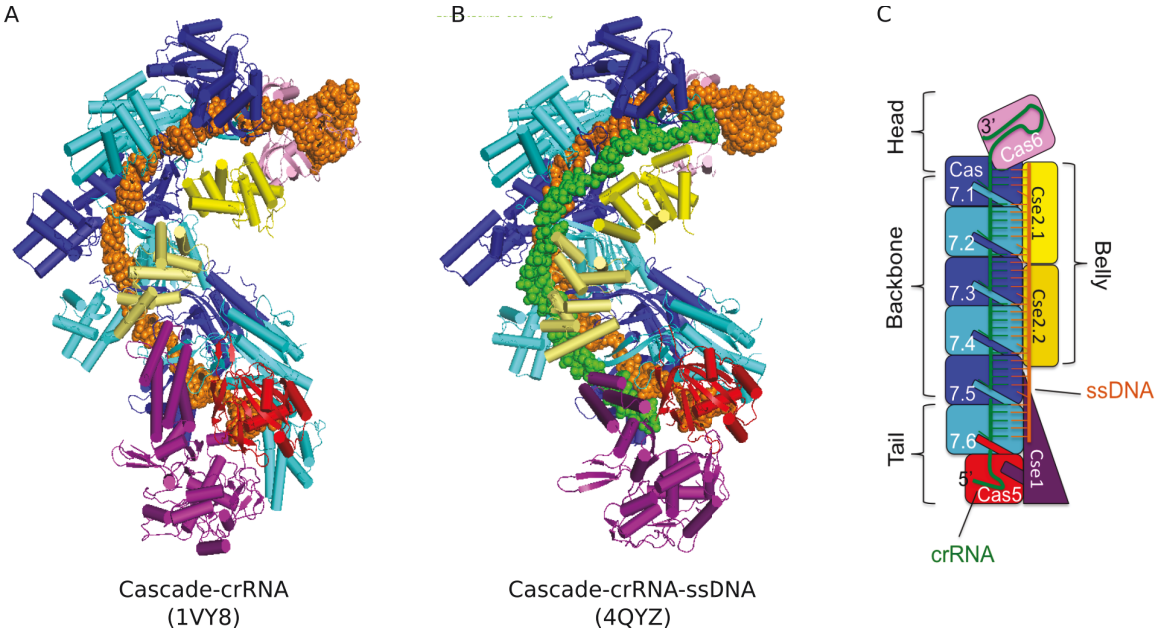

Fig 1. 4 Crystal structures of Cascade. (A) Cascade bound to crRNA. (B) Cascade bound to crRNA and a ssDNA. (C) Schematic of (B).

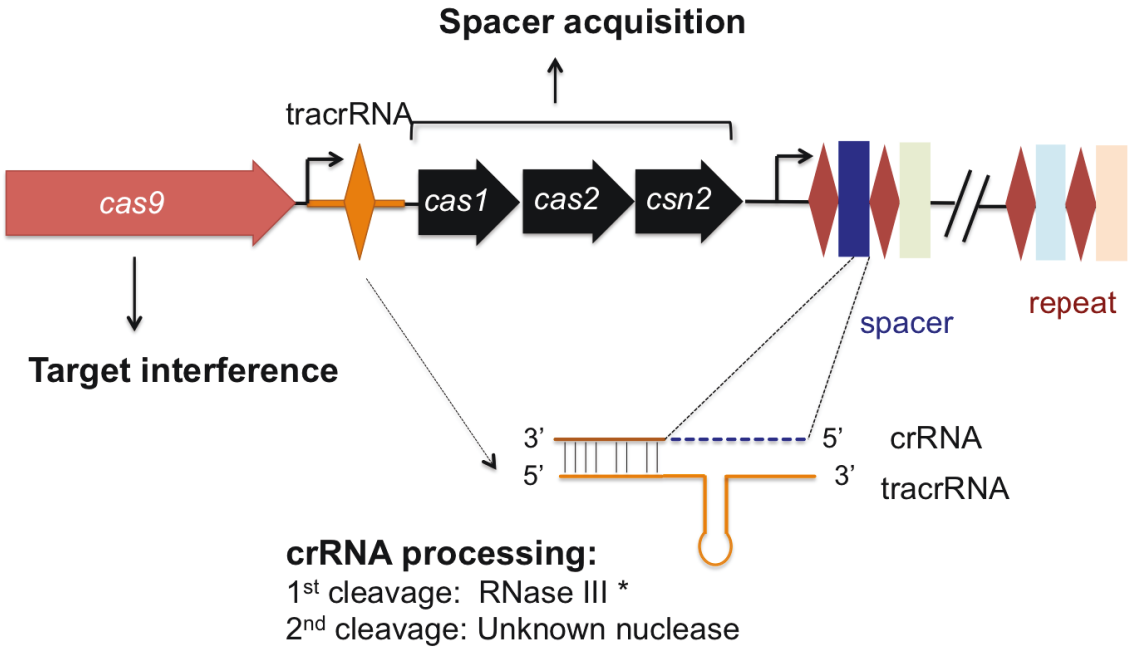

Fig 1. 5 A typical type II CRISPR locus. A *trans*-activating crRNA (tracrRNA) is encoded in a type II cas operon. The orange diamond within tracrRNA gene indicates degenerated repeat sequence. The 1st cleavage event at the repeat complementary region of crRNA and tracrRNA is carried out by RNase III. The asterisk indicates that RNase III is not a Cas protein. The 2nd cleavage event occurred at the 5' end of the crRNA is carried out by an unknown nuclease. The mature crRNA contains 20 nt of the spacer derived sequence (blue dash line).

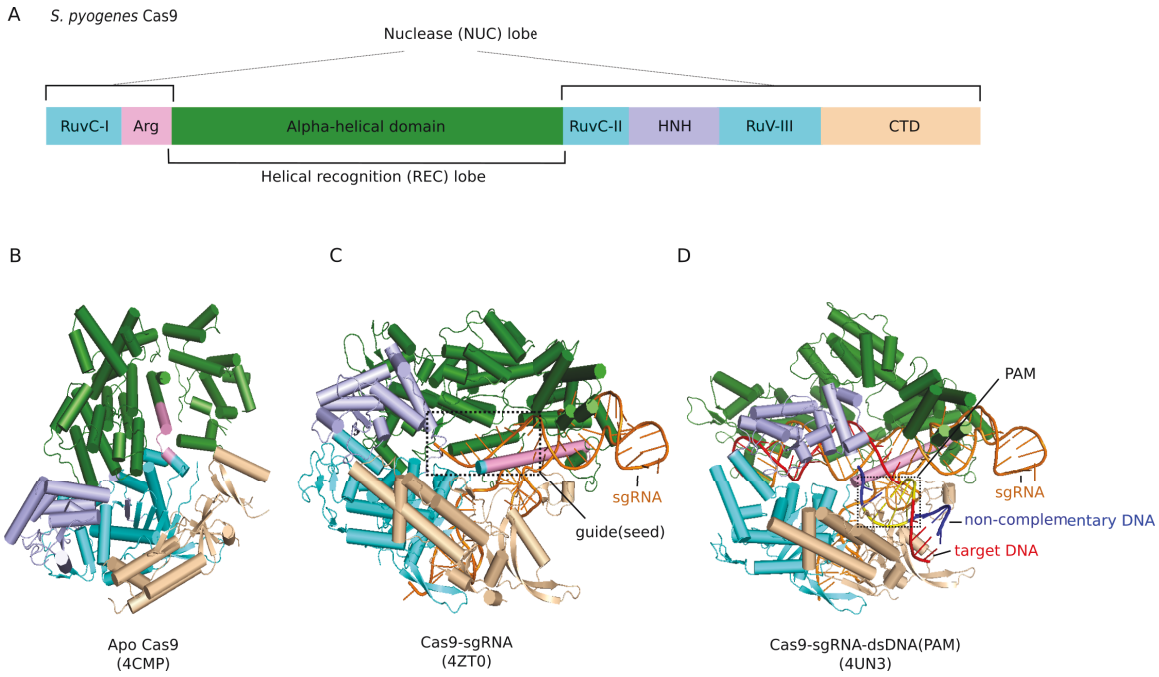

Fig 1. 6 Crystal structures of type II *S. pyogenes* Cas9. (A) Doman organization of *S. pyogenes* Cas9. (B) Crystal structure of apo Cas9. (C) Crystal structure of Cas9 bound to a sgRNA. The dashed box highlights the 10 nt seed region of the guide sequence. (D) Crystal structure of Cas9 bound to a sgRNA and dsDNA containing a PAM. The dashed box highlights the 3 bp PAM (yellow).

Chapter 2

R-loop expansion by Type I *Escherichia coli* Cascade complex

Abstract

In the type I CRISPR-Cas system, a multisubunit ribonucleoprotein complex called Cascade uses its crRNA to detect foreign cognate DNA and recruits a translocating nuclease Cas3 to degrade the DNA. Cascade target identification results in formation of an R-loop, in which the crRNA hybridizes with the target strand and the non-target strand becomes displaced. Here we show that the non-target strand is held by a concave groove on the surface of Cascade, and this binding facilitates DNA strand separation during R-loop formation. Combining structural, biochemical and *in vivo* data, we show that this groove is important for target DNA binding. We demonstrate via single-molecule experiments that efficient R-loop formation is impeded by mutations in the groove. Finally, we show that R-loop stability is enhanced by a locking mechanism, and the locked conformation is sufficient for Cas3 recruitment.

Introduction

Type I systems are the most prevalent CRISPR-Cas systems, accounting for 60% of all CRISPR-Cas systems, and can be further divided into six subtypes (I-A to I-U) (8). The common feature of the type I systems is that they all utilize a structurally similar effector complex termed Cascade (CRISPR-associated complex for antiviral defense) for target recognition and the helicase-nuclease Cas3 for target destruction (16,27,71).

In the type I-E CRISPR-Cas system from *Escherichia coli*, Cascade is a 405-kilodalton (kD) complex comprised of 11 subunits of five Cas proteins ($Cse1_1$, $Cse2_2$, $Cas7_6$, $Cas5_1$, and $Cas6_1$) and a 61-nt crRNA. Target recognition by Cascade requires complementary base pairing between the DNA target and the crRNA spacer sequence, as well as the presence of a 3 bp protospacer adjacent motif—PAM. The first identified PAM for *E. coli* was 5'-CWT-3' (where W is an adenosine or thymidine) (31). Detection of PAM by Cse1 promotes binding of Cascade to the DNA target and enables the formation of an R-loop structure between the crRNA and the dsDNA (33). In the R-loop structure, the target DNA strand hybridizes with the crRNA while the non-target strand becomes displaced. A "locking" step enhances the stability of the R-loop, and is suggested to occur through the conformational rearrangement of Cse1 and Cse2 subunits (20,23,72,73). Finally, the trans-acting helicase-nuclease Cas3 is recruited to the complex to degrade the DNA (27,71).

The interaction between Cascade and the dsDNA target has been elucidated by both structural and biochemical studies. The crystal structure of a pre-target bound Cascade shows that the crRNA spacer region is displayed as six 5-nt fragments in a pseudo-A-form configuration on the concave surface generated by the six interwoven

Cas7 subunits (22,24). Every sixth nucleotide is flipped out by the protruding long β-hairpin from Cas5 and Cas7.1-Cas7.5 subunits (22-24). Consistent with this observation, the crystal structure of Cascade bound to a ssDNA target reveals that the ssDNA forms a non-canonical ribbon-like structure with crRNA, in which base paring is discontinuous at every sixth position (23). The target strand is stabilized primarily through Watson-Crick hydrogen bonding with the crRNA but also by interactions with the Cse1, Cse2 and Cas7 subunits (23). By contrast, how Cascade interacts with the displaced strand is less clear. One low-resolution cryo-EM reconstruction of Cascade bound to a 72-bp dsDNA target suggests that the 5' end of the displaced strand likely loops around Cse1, allowing Cas3 to access its cut site (21). This is supported by previous footprinting experiments, that showed that the 3' end but not 5' end of the displaced strand is protected by Cascade binding (19). These studies help to understand the overall positioning of the displaced strand on Cascade. However, direct evidence for the location of the displaced strand binding site(s) on Cascade is lacking, nor is the mechanism of R-loop formation fully understood.

We identified a concave groove in the ssDNA-bound Cascade structure that could serve as a binding site for 3' end of the displaced strand (23). This groove is surrounded by basic residues from Cse2 and Cas7 subunits (23). Here we present a mutational analysis of the putative binding site together with biochemical, single-molecule, and *in vivo* experiments demonstrating the critical role of the basic residues in the displaced strand binding. We show that mutating these residues affects Cascade binding to dsDNA targets and hinders R-loop formation. These mutations, however, do not affect the overall

R-loop stability and subsequent Cas3 recruitment. Our study provides mechanistic insights into target unwinding and R-loop formation in type I-E CRISPR-Cas systems.

Results

Structural analysis reveals a basic groove for displaced strand

The structure of ssDNA-bound Cascade (4QYZ) reveals a prominent basic groove between Cse2 and Cas7, distinct from the guide RNA-target DNA binding pocket (Fig 2. 1A). This groove is ~14 Å wide and ~17 Å deep, which could easily accommodate a single-stranded nucleic acid. Based on footprinting experiments (19,33), Cascade protects the 3' end of the non-target strand, which maps to a similar position in the structure. Consistent with these observations, the groove is also lined with several conserved basic residues from Cas7 (Lys^{34}, Lys^{299}, and Lys^{301}) and Cse2 (Arg^{53}, Lys^{142}, Arg^{143}, and Arg^{110}).

Since Cse2 subunits undergo a conformation rearrangement upon target binding, we compared the positioning of the Cse2 residues in the pre-target bound structures (IVY8 and 4U7U) and those in the target bound structure. Interestingly, the basic residues are more dispersedly positioned in the pre-target bound state (Fig 2.1D-F). For example, Arg^{53} on both Cse2.1 and Cse2.2 is tilted outwards, making contact with the main chain carbonyl oxygen from Glu^{4} (Fig 2.1F). Arg^{143} on Cse2.1 faces back to interact with Gln^{147} (Fig 2.1E). These interactions are absent in the target bound state, and the side chains turn into the groove space. Additionally, the side chain of Lys^{142} on both Cse2.1 and Cse2.2 rotates ~90°, becoming nearly perpendicular to the surface, poised for interactions. Overall, the basic residues become more structurally aligned in the target bound state, forming a prominent positively charged path on the surface of the Cse2 dimer (Fig 2.1B). Notably, Cse2.1 undergoes a ~10 ° rotation towards the bottom of the

groove, causing a 30° turn in the path. This rotational twist may add an extra grip on the non-target strand (Fig 2.1B).

Based on the structural analysis, we hypothesize that the basic groove binds to the PAM distal end of the non-target strand during R-loop formation.

The basic groove is important for dsDNA binding

To test whether the basic groove plays a role in dsDNA binding, we mutated a subset of the identified basic residues, namely Cse2 R53E, K142E, R110E, and Cas7 K34E, and purified Cascade complexes bearing each point mutation. Because there are two copies of Cse2 and six copies of Cas7, Cse2 mutations display on both copies and Cas7 mutations display on all six copies. The Cascade mutants purify like wild-type (WT), and the subunits harboring the mutations are incorporated at a similar level to WT (Fig 2.2B and C), suggesting that the mutations do not disturb Cascade complex assembly. The co-purified guide RNA was also extracted from each mutant and analyzed on a urea denaturing gel (Fig 2.2D). Only RNA from Cse2 R110E shows slight degradation, suggesting that this mutant may have a minor defect in shielding the guide RNA (Fig 2.2D).

We next examined the binding activity of these Cascade mutants to a dsDNA target containing a PAM and a complementary protospacer sequence. A double-filter binding assay was employed to assess the binding affinity of Cascade to DNA. We incubated a trace amount of radiolabeled dsDNA with increasing concentrations of Cascade, and passed the reactions through nitrocellulose and positively charged nylon membranes to collect protein-bound DNA and free DNA, respectively. 250 nM of Cse1

was supplemented in all cases to prevent dissociation of Cse1 at low concentrations of Cascade, as reported previously (33,74).

We first tested the function of a 5'-CTT-3' versus a 5'-CAT-3'. Interestingly, we found a 5-fold decrease in the apparent dissociation equilibrium constant (K_d) of DNA binding to a 5'-CTT-3' (1.3 ± 0.7 nM) compared to a 5'-CAT-3' (6.1 ± 0.4 nM) PAM (Fig 2.3A). Furthermore, in our single-molecule magnetic tweezer experiments (detailed below), we also observed a ≥ 10-fold difference in the mean time of R-loop formation between a 5'-CTT-3' versus a 5'-CAT-3' PAM at the same Cascade concentration of 90 nM (Fig 2.5D). The mean time is still higher when using 90 nM of Cascade and the 5'-CAT-3' PAM (Fig 2.5D) versus 10 nM of Cascade and the 5'-CTT-3' PAM. Thus, the 5'-CTT-3' PAM was chosen preferentially for this study.

Using the double-filter binding assay, we observed an increase in K_d for all four mutants, with ~20-30 fold for Cse2 R53E, Cse2 K142E and Cas7 K34E, and ~200-fold for Cse2 R110E (Fig 2.3B left). This result suggests that these residues are important for dsDNA binding. To further confirm that these mutations specifically hinder R-loop formation during binding, we created a DNA substrate containing a "bubble" at the protospacer region to mimic a pre-formed R-loop structure. WT Cascade binds to this bubble DNA at a K_d of 0.17 ± 0.02 nM, ~ 8 times tighter than that of the dsDNA (Fig 2.3B right). This corresponds to a difference in Gibbs free energy of -5.2 kJ/mol, which was provided by separating the 32 bp of dsDNA. As expected, all mutants display within experimental error WT level of affinity to the bubble DNA. Together, these data indicate that these residues are likely involved in R-loop formation, presumably by separating the two DNA strands.

Mutations in the groove affect in vivo Cascade-mediated immunity

In order to further evaluate the function of these residues, we set up an *in vivo* CRISPR interference assay using *E. coli* BL21-AI strains. Plasmids encoding Cascade, crRNA and Cas3 were transformed into BL21-AI cells, and their expression was under control of a T7 promoter which could be further fine-tuned by L-Arabinose from 0.001 to 0.2%. After growing in the presence of L-Arabinose and IPTG, the cells were made competent and transformed with target plasmids to assess the CRISPR interference activity. This assay is similar in principle to previously described plaque assays where BL21-AI cells overexpressing type I-E system were challenged with Lambda phages (16). Sashital *et al*. noted that overexpression of this system could allow Cascade to overcome stringent binding defects such as PAM mutations, and attenuating the system by reducing the amount of inducers (0.02% L-Arabinose and 0.01 mM IPTG) enabled distinction between WT interference and defective interference against PAM mutations (33). We presumed that defects in non-target strand binding are less severe than PAM mutations, so, in order to avoid masking of these defects, we further attenuated the system by using 0.002% arabinose and 0.005 mM IPTG.

Using this assay, we assessed the effects of the groove mutations on Cascade-mediated immunity. We first evaluated WT-level immune response. Immunity is reported as the ratio between the colony forming units (CFU) of a control plasmid (containing no protospacer) and that of a target plasmid. Cells expressing WT type I-E system exhibited a 50-fold stronger immunity against a WT target plasmid harboring a 5'-CTT-3' PAM and a complementary protospacer versus a control plasmid containing no protospacer

(Fig 2.4A). Replacing 5'-CTT-3' with a 5'-CAT-3' PAM lowered the immunity (Fig 2.4A), in agreement with our *in vitro* observations (Fig 2.3A and Fig 2.5D). Changing to a 5'-CGT-3' PAM completely abolished the immunity (Fig 2.4A). For cells expressing mutant Cascade, we observed on average a 10-fold reduction in CRISPR immunity against a WT target plasmid, with a pattern mostly mimicking *in vitro* data (Fig 2.4B). Since these mutations do not affect Cascade assembly, we hypothesized that the observed reduction in immunity is due to deficiency in DNA binding and possibly Cas3 recruitment.

Real time R-loop observation reveals defects in R-loop formation but not stability of R-loop

In collaboration with Dr. Ralf Seidel's laboratory (Universität Leipzig, Germany), we employed a single molecule magnetic tweezer technique to observe real time R-loop formation by *E. coli* Cascade. R-loop formation as well as its length and stability can be characterized at the single-molecule level using magnetic tweezers, where the amount of DNA untwisting upon R-loop formation is detected as previously shown for *Streptococcus thermophilus* Cascade and Cas9 (72,73). Briefly, a 2.1k bp DNA molecule containing a protospacer is immobilized on a fluidic cell at one end and attached to a magnetic bead at the other. A pair of magnets placed above the cell stretches the DNA. Depending on the distance between the bead and the magnet, a range of forces could be applied to stretch the DNA to different extents. Moreover, the rotation of the magnets allows the supercoiling of the DNA either negatively or positively. When holding the

DNA at constant negative supercoiling, the formation of an R-loop by Cascade is seen as a change of the DNA length, which is recorded as a function of time (Fig 2.5A).

Dr. Ralf Seidel and colleagues performed magnetic tweezer experiments using purified *E. coli* Cascade provided by us. They first investigated the R-loop formation by Cascade carrying a crRNA fully matching a protospacer sequence containing a 5'-CTT-3' PAM (Fig 2.5B). The DNA is first twisted at a low force to generate negative supercoiling in favor of R-loop formation. In presence of Cascade, we observed a sudden change of DNA length, which is characteristic of an R-loop formation (Fig 2.5B). This change of around 150 nm corresponds to a 2.7 turn's shift of the supercoiling curve, in agreement with the formation of a 32 bp R-loop. Next, the DNA was rewound to a positive supercoiling state to allow the dissociation of Cascade. Cascade remained bound to the DNA at a relative high torque (27 pN nm), confirming the locked state described previously (72). However, even under elevated force and maximum torque (36 pN nm), they were unable to dissociate Cascade from the DNA. In contrast to *S. thermophilus* Cascade, which has a mean time for dissociation of 3 seconds at a positive torque of 22 pN nm (72), *E. coli* Cascade exhibits a much stronger locking mechanism.

Previously, it has been shown that the locking of *S. thermophilus* Cascade is impaired if mismatches are introduced at the PAM distal end of the protospacer. A patch of 4 mismatches decreased the stability of the R-loop by 50% while a 6 mismatches patch completely abolished the locking (72). In order to loosen the locking by *E. coli* Cascade, a DNA substrate containing a protospacer with 6 mismatches at the PAM distal end was used (Fig 2.5C). In this case, R-loop was induced at a lower force and lower negative torque value. The change of 120 nm in the DNA length corresponds to a 1.95 turn's shift

of the supercoiling curve in agreement with the formation of a partial R-loop of 26 bp. Furthermore, by twisting back the DNA to a positive supercoiled state we observed the dissociation of Cascade during the rotation, and it occurred at very minimum positive turns.

By abolishing the locking step, Dr. Ralf Seidel and colleagues were able to remove the Cascade in mild conditions to further statistically study the R-loop formation, recording multiple events. They investigated the mean times for R-loop formation on a "non-lockable" protospacer containing a CTT PAM at different negative torques (Fig 2.5D). In contrast to *S. thermophilus* Cascade, *E. coli* Cascade does not show any torque dependence for R-loop formation, although negative supercoiling is needed to initiate R-loop. They then assessed the mean time for R-loop formation of four Cascade mutants in comparison to the WT at specific negative torque (-7 pN nm) and protein concentration (30 nM) (Fig 2.5B). Although all mutants were able to form unlocked R-loops identical to those observed with the WT, they all showed a mean time for R-loop formation at least doubled, confirming their defects in R-loop formation (Fig 2.5E).

The ability of each mutant to lock the R-loop was also tested using the fully matching protospacer. They found that all Cse2 mutants remained bound as tight as observed previously for the WT (data not shown), whereas the mutation K34E in the Cas7 subunits showed minor deficiency in locking (Fig 2.5F). At the torque of 27 pN nm, they observed a sudden jump of 2.7 turns for Cas7 K34E, indicating the dissolution of the 32 bp R-loop. The mean time for dissociation of Cas7 K34E mutant was around 500 seconds at 27 pN nm torque, which is still significantly stronger in comparison to *S.*

thermophilus Cascade. However, it is not clear in what extend this point mutation of Cascade backbone is impairing the locking stability.

Mutations in the basic groove do not affect Cas3 recruitment

The downstream event after Cascade target recognition is to recruit Cas3 helicase-nuclease for target destruction. We next asked if these mutations affect Cascade's ability to recruit Cas3. We performed a previously described *in vitro* Cas3 cleavage assay (27). We cloned a target plasmid bearing a complementary protospacer and a 5'-CTT-3' PAM. We linearized the plasmid target with the restriction enzyme KpnI, which cuts ~2 kb upstream and ~3 kb downstream of the protospacer. It was previously determined that Cas3 translocates and cleaves DNA preferentially upstream of the protospacer (27,71). Thus, the cleavage product of this substrate by Cas3 is primarily a ~3 kb band (Fig 2.5).

Firstly, we performed Cas3 cleavage assay at 20 nM of Cascade. We incubated Cascade, Cas3 and the target plasmid in the presence of ATP, Mg^{2+} and Co^{2+}. Following incubation, proteins were removed by phenol-chloroform extraction, and the DNA was separated on agarose gel and visualized by ethidium bromide staining. Based on the K_d values from the double-filter binding assay (Fig 2.3B), only WT is fully bound to the target at this concentration. As expected, we observed only robust DNA cleavage for WT, but nearly undetectable levels of cleavage for all mutants (Fig 2.5A). This suggests that Cas3 recruitment is limited by Cascade binding to the target under this condition. Next, we increased Cascade concentration to 1 μM and repeated the Cas3 cleavage assay. At this concentration, DNA binding is no longer a rate-limiting step for all mutants. As a result, robust DNA cleavage was observed for all mutants, suggesting that these

mutations do not affect Cas3 recruitment once they are bound to the DNA (Fig 2.5B). This is consistent with magnetic tweezer data that R-loop stability is barely affected by these mutations.

Discussion

Target recognition is a critical step in CRISPR mediated immunity. In type I systems, Cascade is responsible for identifying foreign targets and recruiting the translocating Cas3 nuclease. Successful target recognition results in an R-loop formation between the crRNA and dsDNA. While the target strand base pairs with the crRNA forming a ribbon-like structure (23), the non-target strand become single-stranded, which is a key pre-requisite for Cas3 loading. However, the mechanism for displacing the non-target strand is poorly understood, nor is the path of the non-target strand clear. The cryoEM structure of Cascade bound to a dsDNA target suggests the PAM proximal end of the non-target strand possibly loops around the four-helix bundle of Cse1 (21). The crystal structure of ssDNA-bound Cascade reveals a possible groove for the PAM distal end of the non-target strand. In this study, we provide experimental evidence that this groove serves as the docking site for the PAM distal end of the non-target strand. The data presented here helps to elucidate a detailed mechanism of R-loop formation by Cascade.

Recognition of a PAM motif initiates target binding. *E. coli* CRISPR PAM was first identified by Mojica *et al*. (31) as 5'-CWT-3'. Subsequent studies have demonstrated that both CTT and CAT lead to efficient interference *in vitro* and *in vivo* (21,25,27,75). However, the functionality of a 5'-CTT-3' PAM versus 5'-CAT-3' in *E. coli* CRISPR interference has not been compared experimentally. Our data strongly suggests that *E. coli* K12 Cascade recognizes a CTT PAM more efficiently than a CAT (Fig 2.3A, 2.5A and 2.5D). Interestingly, bioinformatics analysis of *E. coli* spacer precursors shows a strong bias to 5'-CTT-3' when the CRISPR repeats end in CTC, and to 5'-CAT-3' when

the repeats end in CAC (31). In *E. coli* K12 strain, the repeat sequence is CGGTTTATCCCCGCTGGCGC-GGGGAACTC, which suggests that CTT is likely to be favored. However, the molecular mechanism of PAM recognition by Cascade still remains elusive. A L1 motif from Cse1 is suggested to be involved in PAM interaction (33). Mutating F129 and N131 from the L1 motif significantly weakens PAM protection on the target strand (33). A crystal structure of Cascade bound to a PAM will be desirable for understanding the preference for certain base-pairs at each position.

After PAM recognition, Cascade unwinds DNA in a unidirectional manner (72,73). Interaction with PAM destabilizes the adjacent DNA helix, allowing crRNA to probe the seed sequence (position 1-8) (76,77). Molecular dynamics simulation model of Cascade bound to dsDNA indicates that a β-hairpin on Cse1 may be involved in immediate strand separation after PAM (78). As crRNA continues to base pair with the target strand, more non-target strand become displaced. Our data suggests that further strand separation is likely facilitated by "trapping" the distal end of the non-target strand into the deep groove between Cse2 and Cas7 subunits. Our double-filter binding assay shows that mutating the basic residues in the groove results in a ~20-200 fold deficiency in binding to dsDNA, but nearly no effects on binding to a bubble DNA (Fig 2.3B). The overrepresentation of basic residues in the groove may explain the observed moderate defects even when two or six copies of mutations were introduced. Our data is also in agreement with previous footprinting data that shows the PAM distal end of non-target strand is protected by Cascade (19,33). Stabilizing the non-target strand at a distinct binding site away from the crRNA-DNA duplex prevents the re-hybridization of the dsDNA and could contribute to the stability of the R-loop structure.

Target DNA binding triggers a concerted conformational change of the Cse1 and Cse2 subunits in Cascade, which further induces a locking mechanism that is required for Cas3 recruitment (20,23,72). Both cryoEM and crystal structure show that upon target binding the two Cse2 subunits slide down ~16 Å towards the tail and Cse1 undergoes a ~30° rotation (20,23). Because the groove mutants bind to the bubble DNA with WT affinity (Fig 2.3B), we speculate this conformational movement is not impaired by the mutations. As expected, the magnetic tweezer data reveals the mutants only hinder R-loop formation by kinetic inhibition, but the R-loop stability is not compromised once formed (Fig 2.5). This locked conformation appears to be sufficient to recruit Cas3, despite the deficiency in non-target strand binding (Fig 2.6). It was previously shown that mutations in protospacer stall R-loop formation; however, once the R-loop bypasses the mutation, it proceeds to protospacer end and becomes locked, and Cas3 cleaves the target like WT, regardless of the mutations (73). Our data supports the bidirectional ‘‘telecommunication’’ model proposed by Rutkauskas *et al*. (73) in which the conformational lock by Cse2 rearrangement and additional PAM verification by Cse1 trigger target destruction by Cas3.

Taken together, our study adds to current understanding of R-loop formation by Cascade. Our data suggest that after initial PAM recognition and seed hybridization, Cascade uses two distinct binding pockets for target strand and non-target strand binding (Fig 2.7). Once the R-loop proceeds towards the PAM distal end of the protospacer, Cascade locks the conformation and recruits Cas3 for degradation. The locking mechanism, also observed for *S. thermophilus* Cascade, defines a unique feature for the type I-E systems. Our data presents the first direct observation of R-loop formation of *E.*

coli Cascade by the single-molecule magnetic tweezer technique. Consistent with previous observations (27,71), our data suggests that *E. coli* and *S. thermophilus* Cascade complexes function analogously. Both complexes appear to use the same directional R-loop propagation and locking mechanism to achieve target recognition. Notably, *E. coli* Cascade R-loop formation is torque-independent, and the locking appears to be significantly stronger. Due to a lack of structural information of *S. thermophilus* Cascade, it is unclear what contributes to mechanistic differences. Nevertheless, we think the mechanism proposed here is likely to be conserved in type I-E systems.

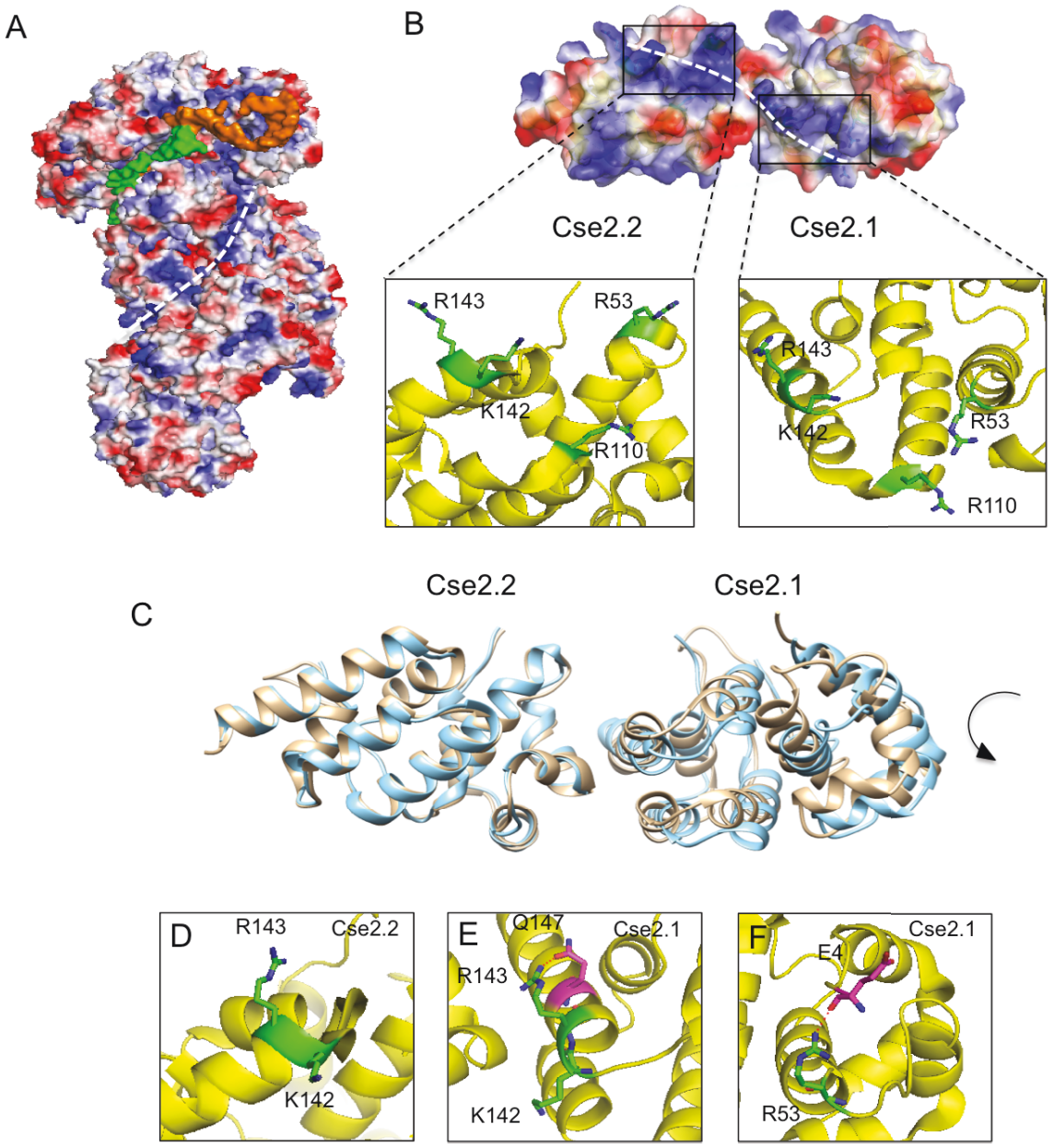

Fig 2. 1 Structural analysis reveals a basic groove for the non-target strand. (A) Electrostatic surface representation of Cascade bound to crRNA (orange) and crRNA (green) illustrates the basic groove (dashed line) for non-target strand binding. (B) Electrostatic surface representation of Cse2 dimer with close-up views of the positively charged residues on each monomer. (C) Overlay of the Cse2 dimer before (gold) and after (cyan) binding to a DNA target. (D-F) Close-up views of the identified residues in the pre-target bound state.

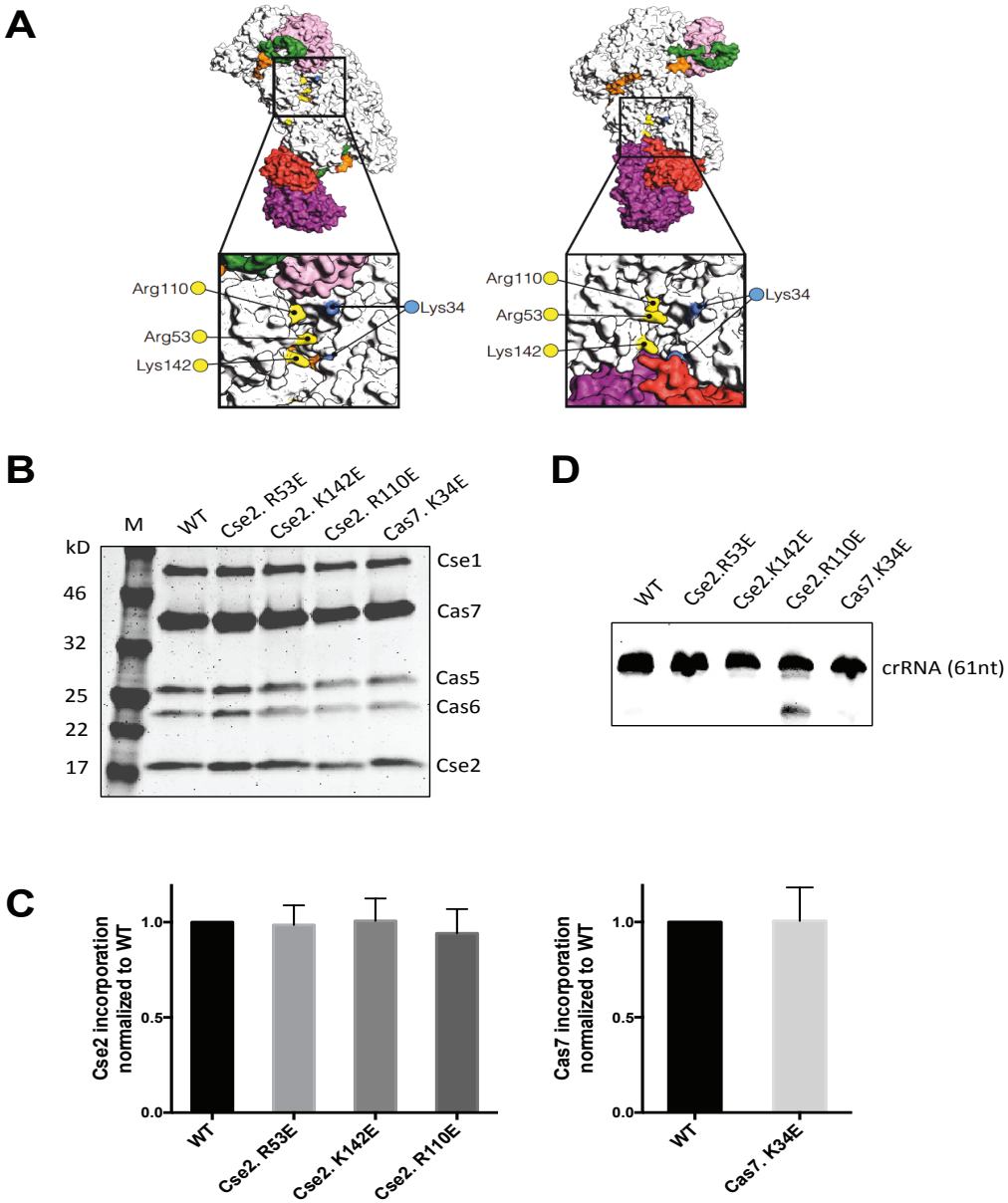

Fig 2. 2 Selected point mutations do not interfere with complex formation. (A) Positions of the selected point mutations in the Cascade structure. Cse2 residues are indicated in yellow, and Cas7 residues are indicated in blue. (B) SDS-PAGE of purified WT and mutant Cascade. (C) Incorporation of Cse2 (left) and Cas7 (right) subunits for the Cascade mutants. The amount of mutated subunit relative to Cas5 was determined from SDS-PAGE band intensities and the ratio was normalized against that of WT. Aggregate data from three replicates are shown, with error bars representing one standard deviation. (D) Denaturing polyacrylamide gel of crRNA extracted from WT and mutant Cascade.

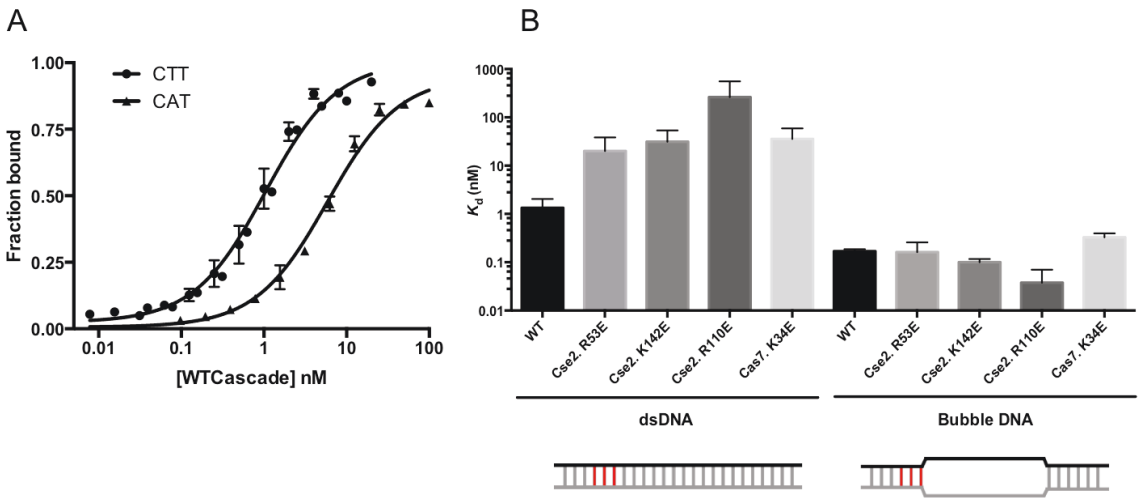

Fig 2. 3 The basic groove is important for dsDNA binding. (A) WT Cascade binding curves measured by double filter binding assay. Binding to a target containing a CTT PAM is shown in dots, and a CAT PAM in triangles. (B) Bar graph plotting K_d values of WT and mutant Cascade binding to a dsDNA substrate (left) and a bubble DNA substrate (right), and schematic drawing of the two substrates. Both substrates contain CTT PAM, indicated as red bars in the schematic. The bubble DNA was designed such that the protospacer region of the non-target strand could not base pair with the target strand. Aggregate data from three replicates are shown, with error bars representing one standard deviation.

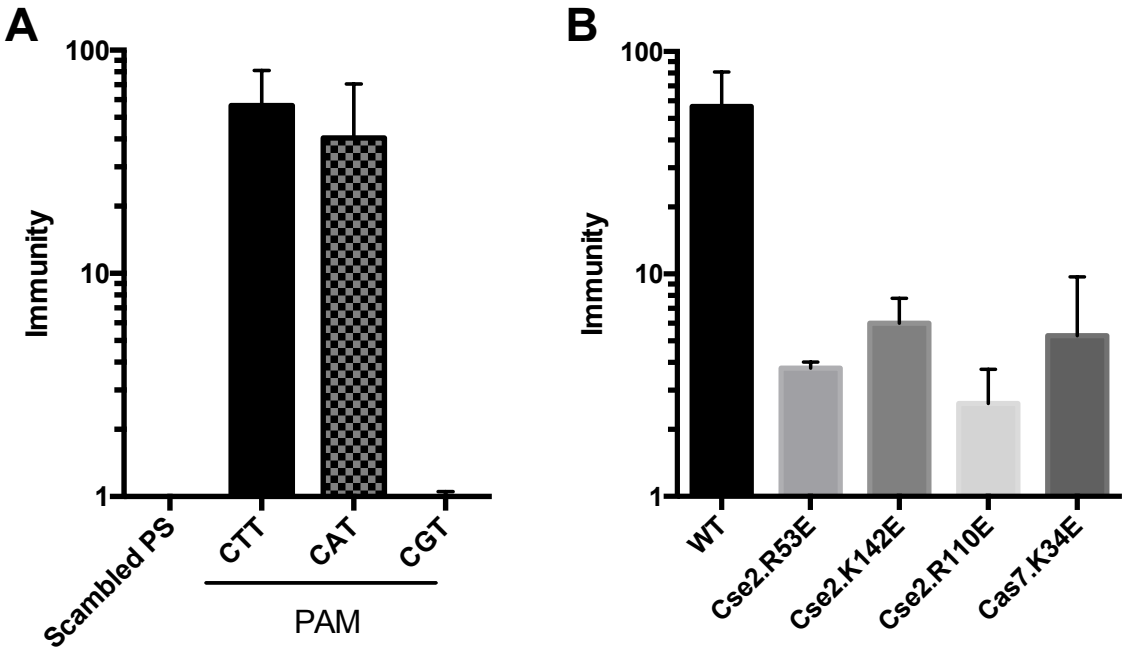

Fig 2. 4 *in vivo* plasmid challenge assay. (A) Relative immunity of *E. coli* cells expressing WT type I-E system against different target DNA. (B) Relative immunity of *E. coli* cells expressing type I-E system with WT or mutant Cascade against a CTT target. In both (A) and (B), immunity is calculated as the ratio between the colony forming units (CFU)/ug of a control plasmid (scrambled protospacer) versus that of a target plasmid in a given trial. Aggregate data from three replicates are shown, with error bars representing one standard deviation.

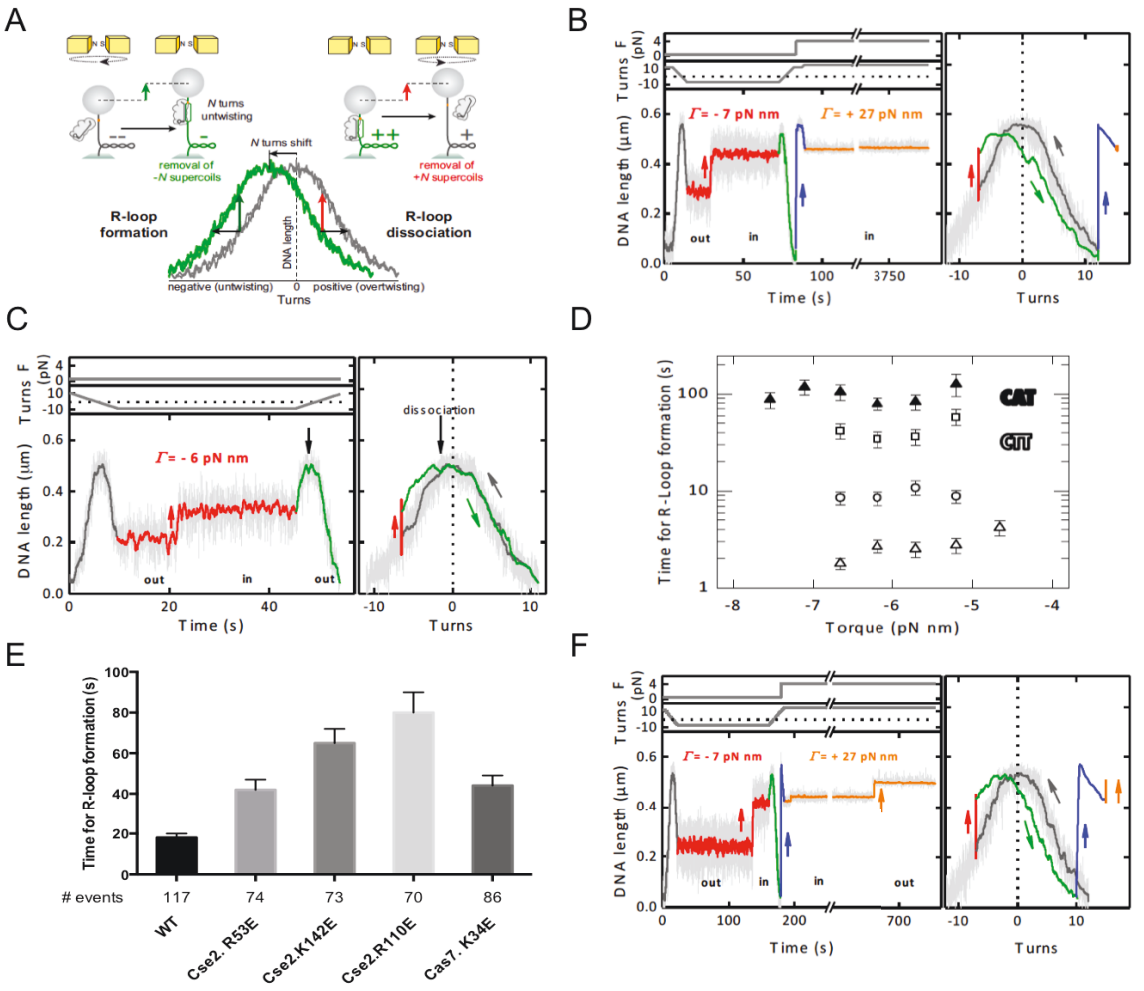

Fig 2. 5 Real time R-loop observation using single molecule magnetic tweezer technique. (A) Principle of R-loop detection through single-molecule supercoiling (adapted from (73)). DNA molecules are attached to magnetic beads and magnet rotation changes DNA supercoiling. R-loop formation causes local DNA unwinding and thus overwinding of the adjacent DNA. The grey curve is DNA length plotted against introduced turns in absence of an R-loop. The plot is shifted to the left by the number of helical turns being unwound by the R-loop (green curve). When DNA is under constant negative supercoiling (favorable for R-loop formation), R-loop formation is indicated as a DNA length change (green arrow). Similarly, R-loop dissociation is observed at positive supercoiling as a similar length change in DNA (red arrow). (B) R-loop formation-dissociation cycle for a fully matching protospacer with CTT PAM. R-loop formation was seen at -7 pN nM (red arrow) but no dissociation was observed even at high positive torque (constant orange line). (C) R-loop formation-dissociation cycle for a protospacer containing a 6-nt mismatch patch at the PAM distal end. Cascade release is seen when DNA has no negative torque (green line). (D) Time for R-loop formation at different torques for DNA targets containing CTT or CAT PAM. Open symbols, CTT PAM; filled symbols, CAT PAM. Shape of symbols indicates different Cascade concentration used: triangles, 90nM; circles, 30nM, squares, 10nM. (E) Time for R-loop formation by

Cascade mutants. (F) The locked mutant Cas7.K34E dissociates from DNA at high positive torques.

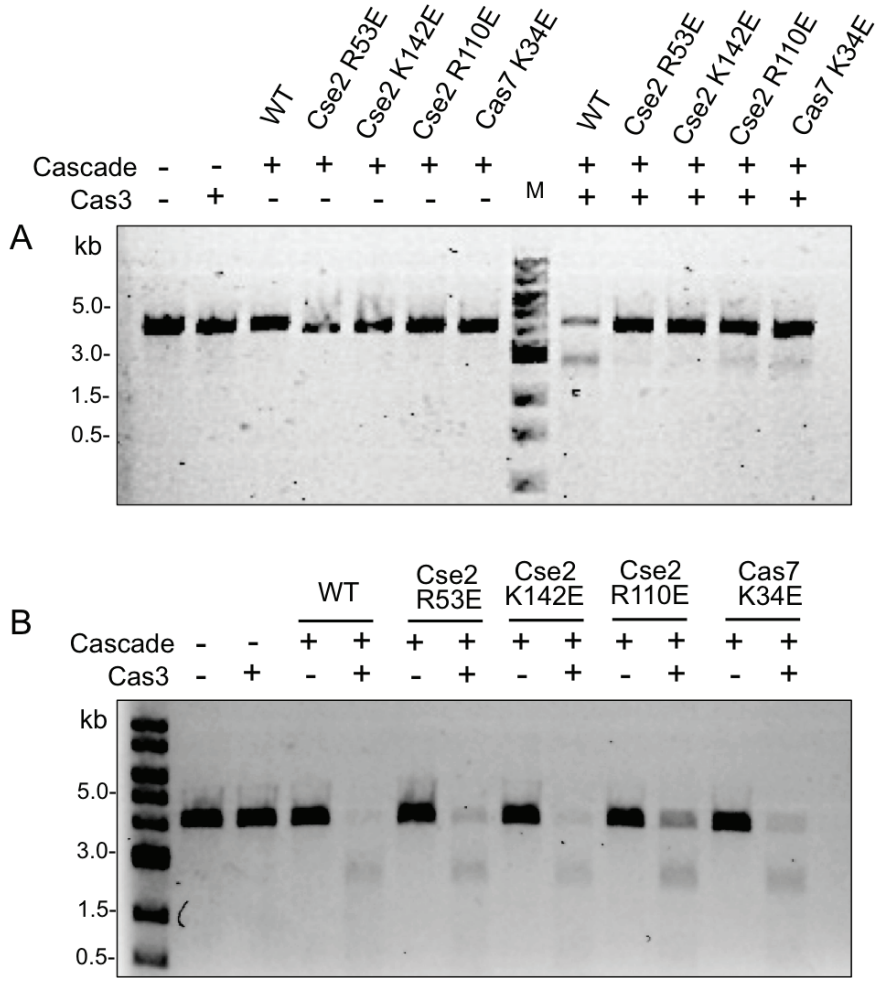

Fig 2. 6 Cas3 recruitment assay. Agarose gels of Cas3 recruitment assay with 20 nM (A) and 1 μM (B) Cascade. Reaction mixtures containing indicated amount of Cascade, 2 nM linear plasmid, 300 nM Cas3, 10 mM Mg^{2+}, 100 μM Co^{2+}, and 2 mM ATP were incubated for 30 min at 37 °C, followed by phenol chloroform extraction before loaded onto the gels.

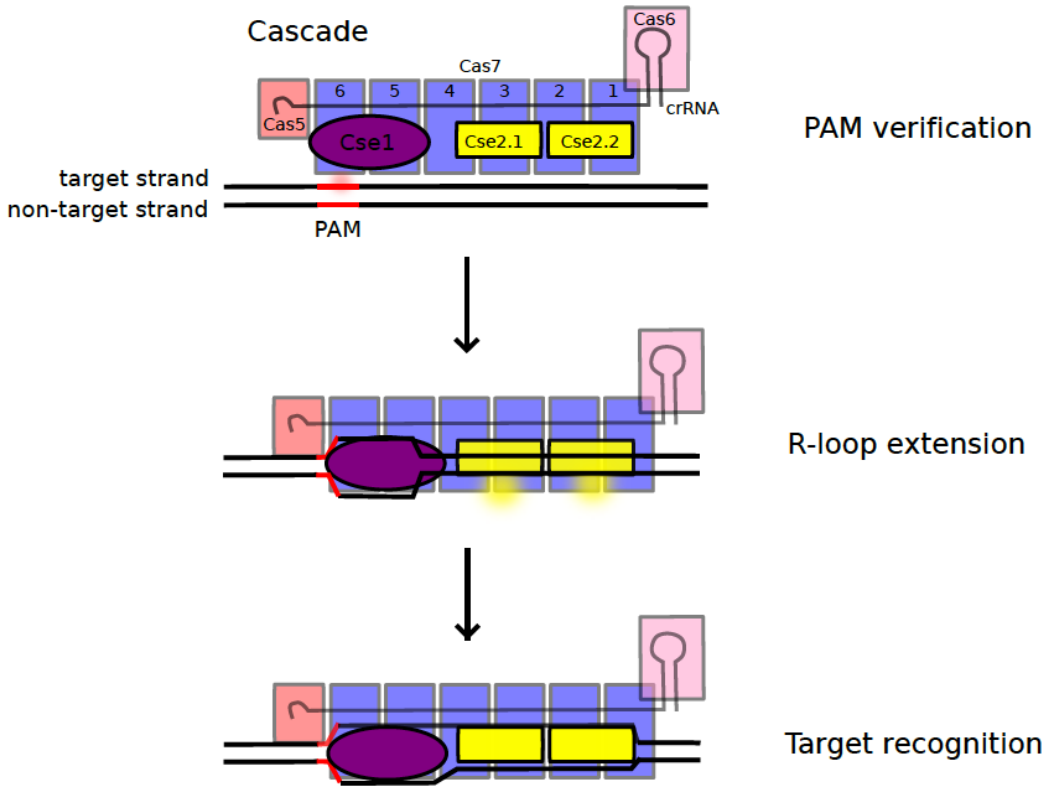

Fig 2. 7 Schematics of R-loop formation by Cascade.

Materials and Methods

Cloning and Mutagenesis

For protein purification, a pBCDE construct containing *E. coli cse2-cas7-cas5-cas7* in the pHAT4 vector, a pCRISPR-A construct containing CRISPR (7x spacer) and *E. coli cse1* in the pRSFDuet-1 vector, and a pCse1 construct containing *E. coli cse1* in the pMAT11 vector were previously generated by Sabin Mulepati. pBCDE and pCRISPR-A were used to express Cascade-crRNA. The CRISPR array consists of seven identical spacers (sequence: 5'-CCAGTGATAAGTGGAATGCCATGTGGGCTGTC-3'). pCse1 was used to express Cse1 needed for double-filter binding experiments. pCas3a containing *E. coli cas3* in a pSAT1 vector was created by John Mallon.

For the plasmid challenge assay, a pABCDE-CR construct containing *E. coli cse1-cse2-cas7-cas5-cas7* and CRISPR (7x spacer) in the pACYCDuet-1 vector, a pCas3b construct containing *E. coli cas3* in the pRSFL vector, and target plasmids in the pBAT4 vector were prepared by Jasvir Kaila. To generated plasmid targets, synthetic oligonucleotides, bearing the appropriate sequence, were annealed and ligated into pBAT4 vectors.

Cascade mutants were created using round-the-horn site-directed mutagenesis. All constructs were verified by DNA sequencing. All plasmids used in these studies are detailed in Table 2.1. Primers and oligonucleotides used are listed in Table 2.2.

Protein expression and purification

Proteins were overexpressed in the T7Express strain of *E. coli* (New England Biolabs). Cells were grown in Luria-Bertani (LB) medium supplemented with the

appropriate antibiotic(s) (Table 2.1) at 37 °C to an OD_{600} of 0.3–0.5, and subsequently protein expression was induced with 0.2 mM isopropyl-β-D-thiogalactopyranoside (IPTG) overnight at 20 °C.

WT Cascade, Cascade mutants, and Cse1 were purified as described previously (74). Briefly, harvested cells were lysed in buffer L (20 mM Tris-HCl, pH 8.0, 100 mM NaCl and 10% glycerol), clarified, and then mixed with 5 ml of profinity immobilized metal affinity chromatography (IMAC) resin (Bio-Rad). The resin was then washed with 10 mM imidazole before the protein of interest was eluted with 250 mM imidazole. Samples were desalted to remove imidazole and then treated with tobacco etch virus (TEV) protease overnight at 4 °C to remove the N-terminal tag. Samples were reapplied to IMAC resin to remove the His_6-tagged TEV protease, any cleaved tag, or any remaining tagged protein. Samples were then concentrated and loaded onto a HiLoad 26/60 S200 size-exclusion column (GE Healthcare) pre-equilibrated with gel filtration buffer (20 mM Tris-HCl, pH 8.0, 200 mM NaCl). Proteins were concentrated to 10 ~ 30 μM, flash-frozen, and stored at -80 °C.

For *E. coli* Cas3 purification, overnight cultures were harvested and immediately lysed in buffer L (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM dithiothreitol (DTT), and 10% glycerol). The cell lysate was clarified by centrifugation, and mixed with 5 ml of IMAC resin. The resin was washed consecutively with buffer L supplemented with 5 mM imidazole and then with 1 M NaCl. The remaining bound proteins were eluted with buffer L supplemented with 250 mM imidazole. The sample was directly loaded onto a HiLoad 26/60 S200 size exclusion column pre-equilibrated in buffer A (20 mM Tris-HCl, pH 8.0, 200 mM NaCl, and 1 mM DTT). Fractions containing Cas3 were pooled and

treated with SENP protease to remove N-terminal His_6-SUMO tag overnight at 4 °C. The cleaved sample was then flowed through IMAC resin to remove the His-tagged SENP protease, any cleaved tag, or any remaining tagged protein. Samples were then concentrated, and loaded onto a HiLoad 26/60 S200 size exclusion column pre-equilibrated with buffer A. Purified Cas3 was concentrated to ~5 μM, flash-frozen, and stored at -80 °C.

crRNA extraction

crRNA was isolated from Cascade using phenol:chloroform:isoamyl alcohol (25:24:1) (Sigma-Aldrich) extraction, followed by ethanol precipitation. RiboLock (Thermo Fisher Scientific) was used at 1 unit/ μL to prevent RNA degradation. Isolated RNA samples were analyzed on a 10% denaturing TBE gel and visualized by staining with SYBR Gold (Invitrogen).

Double-filter binding assay

Binding of Cascade to DNA was assessed using a double-filter binding assay (79). All DNA oligonucleotides were gel-purified. dsDNA or bubble substrates (Table 2.2) were made by annealing each strand and purified on 12% native polyacrylamide gels containing 1× TBE. Trace amounts (10-200 pM) of 5'-end ^{32}P-labled DNA targets were incubated with increasing concentrations of Cascade (0-2 μM) for 1 h at 37 °C. Binding reactions contained 20 mM Tris-HCl pH 8.0, 100 mM NaCl, 500 nM competitor (Table 2.2), 0.1 mg/ml bovine serum albumin (BSA), and 10% glycerol. A fixed concentration (250 nM) of Cse1 was supplemented to prevent dissociation of Cse1 at low

concentrations of Cascade, as determined previously (74). Nitrocellulose (LI-COR) and Hybond-N+ nylon (Amersham) membranes were soaked in reaction buffer and assembled onto a 10-well (1-inch diameter) vacuum manifold in the order of gel blot paper, nylon, and nitrocellulose. After prewashing with 1 ml buffer, 100 µl reaction mixture were applied by vacuum, followed by 1 mL of buffer to wash out unbound samples. Filter membranes were dried and counted for radioactivity with a scintillation counter (Beckman). Data analysis was performed with GraphPad Prism software. Reported K_d values are the average of at least three replicates.

*Single-Molecule Experiments**

Single-molecule assays with *E. coli* Cascade were performed in 20 mM Tris·HCl pH 8.0, 100 mM NaCl, and 0.1 mg/mL BSA. Measurements were performed using 30 nM Cascade unless otherwise indicated. After DNA stretching and initial characterization of the DNA, proteins were added, and changes in DNA length were recorded as a function of applied force and DNA turns.

*Cas3 cleavage assay**

Cas3 cleavage assays were performed similarly as previously described (27). Briefly, the plasmid target containing a 5'-CTT-3' PAM was linearized using the restriction enzyme KpnI. 1 nM linearized plasmid, indicated amounts of Cascade, 300 mM Cas3, and 2 mM ATP were mixed in a reaction buffer containing 5 mM HEPES pH 7.5, 60 mM KCl, 10 mM $MgCl_2$, and 100 µM $CoCl_2$. The reactions were incubated at 37 °C for 30 min and terminated by addition of 20 mM EDTA. The proteins were removed

by phenol:chloroform:isoamyl alcohol extraction, and the DNA was separated on 1% agarose gels, stained with ethidium bromide, and visualized using FLA-7000 (Fuji).

*In vivo transformation assay**

The recipient strain was generated by co-transforming *E. coli* BL21-AI cells with pABCDE-CR and pCas3b plasmids (Table 2.1). The next day, 2-3 colonies were picked and grown in LB medium (supplemented with 0.002% arabinose, 0.005 mM IPTG, as well as chloramphenicol and kanamycin) at 37 °C until the OD_{600} was ~ 0.3. A 5 ml culture was used for each plasmid transformation assay. Cells were made competent using a previously described $CaCl_2$ heat-shock procedure (99). Cells were pelleted (2,500 rpm for 10 min) and washed with 3 ml of "Na solution" consisting of 5 mM Tris-HCl pH8.0, 100 mM NaCl, and 5 mM $MgCl_2$. Cells were resuspended in 3 ml "Ca solution" consisting of 5 mM Tris-HCl pH8.0, 100 mM $CaCl_2$, and 5 mM $MgCl_2$ and incubated on ice for 20 min. Cells were pelleted again and resuspended in 200 uL "Ca solution". Cells were then transformed with 80 ng of target plasmid and incubated on ice for another 20 min. After 45 seconds at 42 °C, cells were added with 800 uL LB and incubated at 37 °C for 45 min. Cells were plated with appropriate dilution factors on selection plates. Total transformation efficiency was assessed with a plasmid containing scrambled protospacer in each assay. Immunity was defined as $CFU_{\text{control plasmid}}/CFU_{\text{target plasmid}}$. Reported values are the average of at least three replicates.

* The single-molecule experiments were performed by Christophe Rouillon, a post-doc fellow from Dr. Ralf Seidel's laboratory (Universität Leipzig, Germany); Jasvir Kaila, a ScM student (2015-2016) worked on the *in vivo* assay setup, and performed the *in vivo* experiments with different PAM targets; John Mallon, a doctoral student in our laboratory, generated reagents for the in vitro Cas3 recruitment assay and helped with assay setup.

Table 2. 1 Plasmids used in these studies

Clone	Vector	Gene/sequence
pBCDE[1]	pHAT4 (Amp^r)	*E. coli cse2-cas7-cas5-cas6*
pCRISPR-A[1]	pRSFDuet-1 (Kan^r)	MCS1: *E. coli cse1*; MCS2: CRISPR (7x spacer)
pCse1	pMAT11 (Amp^r)	*E. coli cse1*
pCas3a[2]	pSAT1 (Amp^r)	*E. coli cas3*
pABCDE-CR[3]	pACYCDuet-1 (Cam^r)	MCS1: *E. coli cse1- cse2-cas7-cas5-cas6*; MCS2: CRISPR (7x spacer)
pCas3b[3]	pRSFL (Kan^r)	*E. coli cas3*
pTarget[3]	pBAT4 (Amp^r)	PAM+protospacer

1: pBCDE and pCRISPR-A are used to express Cascade

2: pCas3a was used to express Cas3

3: pABCDE-CR, pCas3b, and pTarget are used in plasmid challenge assays

Table 2. 2 Primers and Oligonucleotides used in these studies

	Sequence (5' to 3')
Primers for round-the-horn mutagenesis	
BR53E forward	AACACCAGCAGGCTCTTTTGCGC
BR53E reverse	CTGGGTTTTCCCAACCAAAAGG
BK142E forward	ACGCGAACGCCAGCAACTTCTG
BR142E reverse	TCTCCCCACCAGGTCAACATCC
BR110E forward	AAACAGCCGATATGGTCCAGTTAC
BR110E reverse	CGTCAGCCCGAATTAATTGAAAG
CR34E forward	AAAGACGAGTAAGAATTTCAAG
CR34E reverse	CGCCGCCGAAAATAGCGTCTTTC
Oligonucleotides used in double-filter binding assay	
T top	AGCGACTCCCGAGCAATCAGACAGCCCACATGGCATTCCACTTAT CACTGGCTTGCTTTCGGCTTGCCGCGC
T bottom	GCGCGGCAAGCCGAAAGCAAGCCAGTGATAAGTGGAATGCCATG TGGGCTGTCTGATTGCTCGGGAGTCGCT
T (bubble)[1]	GCGCGGCAAGCCGAAAGCAAGCTGTCGGGTGTACCGTAAGGTGA
bottom	ATAGTGACCTGATTGCTCGGGAGTCGCT
Competitor top	AGCGACTCCCGAGCAATCACTGTCGGGTGTACCGTAAGGTGAAT AGTGACCCTTGCTTTCGGCTTGCCGCGC
Competitor bottom	GCGCGGCAAGCCGAAAGCAAGGGTCACTATTCACCTTACGGTAC ACCCGACAGTGATTGCTCGGGAGTCGCT
Oligonucleotides used to construct plasmid targets	
T forward	CATGGACAGCCCACATGGCATTCCACTTATCACTGGCTT
T reverse	TCGAAAGCCAGTGATAAGTGGAATGCCATGTGGGCTGTC
T (ctrl) forward	CATGAGTGATTTGTGCAATGCCTTGTCCGCTGTCAACTT
T (ctrl) reverse	TCGAAAGTTGACAGCGGACAAGGCATTGCACAAATCACT

1: The bubble substrate was generated by annealing T top and T (bubble) bottom.

Chapter 3

Characterization of a Type II Cas9 from *Streptococcus thermophilus* LMG18311

* Previously published as

Chen, H., Choi, J. & Bailey, S., 2014.
Cut site selection by the two nuclease domains of the Cas9 RNA-guided endonuclease.
The Journal of Biological Chemistry, 289(19), pp.13284–13294.

Abstract

Cas9, the RNA-guided DNA endonuclease from type II CRISPR-Cas system, has been adapted for genome editing and gene regulation in multiple model organisms. In this chapter, we characterize a Cas9 ortholog from *Streptococcus thermophilus* LMG18311 (LMG18311 Cas9). *In vitro* reconstitution of this system confirms LMG18311 Cas9 together with a *trans*-activating RNA (tracrRNA) and a CRISPR RNA (crRNA) cleave dsDNA with a specificity dictated by the sequence of the crRNA. Cleavage requires not only complementarity between crRNA and target but also the presence of a short motif PAM. Here we show that both the efficiency of DNA target cleavage and the location of the cleavage sites vary based on the position of the PAM sequence.

Introduction

A promising tool for genome manipulation (39-41,80-90) and regulation (42,91-93) in a wide variety of organisms has recently been identified in the RNA-guided DNA endonuclease activity of the type II CRISPR-Cas systems. Programmed DNA cleavage requires the fewest components in the type II CRISPR-Cas system, requiring only crRNA, tracrRNA and the Cas9 endonuclease (37,38), the signature gene of the type II system. The system can be further simplified by fusing the mature crRNA and tracrRNA into a single guide RNA (sgRNA) (37). In addition to its role in target cleavage, tracrRNA also mediates crRNA maturation by forming RNA hybrids with primary crRNA transcripts, leading to co-processing of both RNAs by endogenous RNase III (36). Cas9 contains two nuclease domains that together generate a double-strand break in target DNA. The HNH nuclease domain cleaves the complementary strand and the RuvC-like nuclease domain cleaves the non-complementary strand (37,38).

A short signature sequence, named the protospacer adjacent motif or PAM, is characteristic of the invading DNA targeted by the type I and type II CRISPR-Cas systems. The PAM serves two functions. It has been linked to the acquisition of new spacer sequences and it is necessary for the subsequent recognition and silencing of target DNA, reviewed in (94). The sequence, length and position of the PAM vary depending on the CRISPR-Cas type and organism. PAMs from type II systems are located downstream of the protospacer and contain 2 to 5 bps of conserved sequence. A variable sequence, of up to 4 bps, separates the conserved sequence of the PAM from the protospacer. This variable region is often included in the definition of the PAM sequence, but for simplicity, we refer to this variable region as the linker and the conserved

sequence as the PAM. To date, Cas9 from *Streptococcus pyogenes*, *Streptococcus thermophilus* DGCC7710 and *Neisseria meningitidis* have been employed as tools for genome editing or regulation. For these Cas9 orthologs the PAMs are GG, GGNG and GATT, and the linkers are 1, 1 and 4 bps, respectively (37,95,96).

The simplicity of sgRNA design and sequence specific targeting means the RNA-guided Cas9 machinery has great potential for programmable genome engineering. Cas9 can be employed to generate mutations in cells by introducing dsDNA breaks. The capabilities of Cas9 can be expanded to various genome engineering purposes, such as transcription repression or activation, with its nickase (generated by inactivating one of its two nuclease domains) or nuclease null variants (42,92,93,97). Another appealing possibility for the Cas9 system is to target different Cas9-mediated activities to multiple target sites, for example transcriptional repression of one gene but activation of another (98). To achieve this, multiple Cas9 orthologs will need to be employed, as a single ortholog cannot concurrently mediate different activities at multiple sites (98). Therefore to broaden our understanding of Cas9 proteins, we have characterized the Cas9 ortholog from *Streptococcus thermophilus* LMG18311, which we refer to as LMG18311 Cas9. We choose to investigate Cas9 from this organism not only to increase the repertoire of Cas9 orthologs but also because it utilizes a PAM distinct from those previously characterized and its small gene size is compatible with the standard viral vectors used for delivery into exogenous systems *in vivo* (98).

Here we demonstrate that requirements for DNA cleavage *in vitro* and *in vivo* by LMG18311 Cas9 are the same as other Cas9 orthologs. We also reveal the sequence and linker length requirements of the PAM for LMG18311 Cas9. Finally, we show that the

HNH and RuvC-like nuclease domains of Cas9 select the location of their cleavage sites via different mechanisms. The HNH domain catalyzes cleavage of the complementary strand at a fixed position, whereas the RuvC-like domain catalyzes cleavage of the non-complementary strand using a ruler mechanism.

Results

Identifying the PAM for LMG18311 Cas9

The genome of *S. thermophilus* LMG18311 contains two CRISPR-Cas systems, of type II-A and III-A, each associated with a CRISPR loci: CRISPR-1 and CRISPR-2, respectively. The first study of PAM sequences identified a putative PAM for *S. thermophilus* as RYAAA (where R is a purine and Y a Pyrimidine) (4). This sequence was found in natural target sequences matching 41 spacers collected from 13 different *S. thermophilus* strains, including LMG18311. Subsequent studies showed PAM sequences vary greatly, even between different strains, reviewed in (94). Therefore to confirm the PAM sequence for LMG18311 Cas9, we performed BLAST searches to identify potential protospacers in viral and plasmid genomes that matched any of the 33 spacer sequences from CRISPR-1. This search generated 41 unique target sequences, from the genomes of bacteriophage known to infect *S. thermophilus*. We then aligned 50-nucleotide segments from the identified target genomes, inclusive of the 30-nucleotide protospacer and 10-nucleotide flanking regions (Fig 3.1B). In agreement with the previous study (4), inspection of this alignment clearly identified a 5 bp PAM with a consensus sequence, GYAAA, invariantly located 2 bps downstream of the protospacer (Fig 3. 1B and C). The most commonly observed PAM sequence, found in 7 or the 41 target sequences, was GCAAA.

To confirm the identified PAM was functional we used a previously described transformation assay in which *E. coli* cells containing an exogenous type II CRISPR-Cas system are resistant to plasmid transformation, while cells lacking the system are competent for transformation (99,100) (Fig 3.2A). To generate cells containing the type

II CRISPR-Cas system ($CRISPR^+$ cells), compatible vectors encoding either LMG18311 Cas9 or its cognate sgRNA, engineered to contain a 20-nucleotide sequence derived from the first spacer of CRISPR-1 (Fig. 1C), were co-transformed into *E. coli* BL21 (DE3). In this overexpression system the Cas9 and sgRNA genes are under the control of an IPTG inducible T7 promoter. Control cells lacking the CRISPR-Cas system ($CRISPR^-$ cells) were generated by co-transforming compatible empty vectors into *E. coli* BL21 (DE3). We constructed a target and two control plasmids. The target plasmid contained protospacer-1 (whose sequence was identical to the first spacer of CRISPR-1), a 2 bp linker and the identified PAM (GCAAA) (Fig 3.1C). The first control plasmid contained only protospacer-1, while the second control plasmid lacked both protospacer-1 and PAM. The target and control plasmids were then tested for CRISPR-Cas silencing by transformation into the $CRISPR^+$ and $CRISPR^-$ strains in the presence of IPTG and the appropriate antibiotics (Fig 3.2A). The control plasmids transformed into both strains with similar efficiency (Fig 3.2B). The target plasmid failed to transform into the $CRISPR^+$ cells but transformed into the $CRISPR^-$ cells with an efficiency comparable to that of the control plasmids (Fig 3.2B). All of the transformation efficiencies were comparable to those previously reported (100). These results indicate that the identified PAM is functional *in vivo* and that the type II CRISPR-Cas system of *S. thermophilus* LMG18311 protects *E. coli* cells from transformation by plasmid DNA.

Both the PAM sequence and linker length are important for plasmid interference

To investigate the PAM sequence requirements for LMG18311 Cas9, we transformed a series of plasmid targets harboring single-nucleotide mutations throughout

the PAM sequence in the CRISPR$^+$ strain (Fig 3.2C). Only the plasmid containing a mutation at the position 1 guanosine (that is, the PAM nucleotide closest to the protospacer) was transformed, albeit with a reduced (~66%) transformation efficiency compared to the intact PAM sequence (Fig 3.2C). Plasmids containing single mutations to any of the other four positions were resistant to transformation (Fig 3.2C). These results indicate that the guanosine at position 1 is important for PAM function but individually the four other positions have little effect on PAM function.

A 2 bp linker separates the protospacer from the PAM for LMG18311 Cas9 (Fig 3.1B and C). To investigate how linker length affects Cas9 activity, we generated plasmid targets with linkers ranging from 0 to 5 bps in length (Fig 3.2D). We then determined the transformation efficiency for these plasmids into the CRISPR$^+$ cells. The CRISPR$^+$ cells were equally resistant to transformation by a plasmid target with either a linker length of 2 or 3 bps (Fig 3.2D). Plasmids with other linker lengths transformed with efficiencies more similar to the control plasmid (Fig 3.2D) suggesting that plasmids with these linkers were able to escape CRISPR-Cas silencing.

In vitro reconstitution recapitulates in vivo activity

To further investigate the requirements of PAM sequence and linker length, we reconstituted the activity of LMG18311 Cas9 *in vitro*. LMG18311 Cas9 was expressed and purified from *E. coli* (Fig 3.1A). A 42-nucleotide tracrRNA mimicking the processed tracrRNA, and a 42-nucleotide crRNA containing sequence derived from the first spacer of CRISPR-1 (Fig 3.1C), were chemically synthesized. Plasmid targets were incubated with Cas9, tracrRNA and crRNA and then analyzed by electrophoresis through agarose

gels and ethidium bromide staining. As observed for other Cas9 orthologs, cleavage of the plasmid target occurred in the presence of Cas9, tracrRNA, crRNA and Mg^{2+} (Fig 3.3A). Cleavage also occurred when an sgRNA was substituted for the tracrRNA and crRNA (Fig 3.3B). As expected, cleavage was dictated by the sequence of the sgRNA (Fig 3.3C). Also, Cas9 variants with active site mutations in either the RuvC-like domain (D9A) or HNH domain (H599A) nicked the plasmid targets, while a variant with a double mutation (D9A, H599A) displayed no activity (Fig 3.3D). Cleavage assays using short oligonucleotide substrates confirmed that the HNH domain cleaves the strand complementary to the guide RNA, while the RuvC-like domain cleaves the non-complementary strand (Fig 3.3E). Mapping the location of the cut sites revealed that, as seen with other Cas9 orthologs (7,37,38,101), cleavage of both strands occurs within the protospacer, 3 bps from its PAM proximal end, producing a blunt-end dsDNA break (Fig 3.3E).

We next wished to confirm that either mutations in the PAM or that changes in linker length had the same effect on DNA interference *in vitro* as they did *in vivo*. Therefore, we monitored cleavage of these variant plasmids by recombinant LMG18311 Cas9. The fraction plasmid cleaved was calculated using the procedure detailed in the Materials and Methods section, which accounts for the different binding affinity of ethidium bromide to linear and supercoiled DNA. Consistent with the *in vivo* results mutation of the guanosine at position 1 had the greatest effect, individual mutations to the other four positions of the PAM had only a modest effect on plasmid cleavage (Fig 3.3F). Cleavage of plasmid targets with different linker lengths was optimal at 2 or 3 bps and then decreased steadily with increasing or decreasing lengths (Fig 3.3G).

Metal dependency of DNA cleavage by Cas9

To evaluate whether other divalent cations besides Mg^{2+} can activate DNA cleavage by Cas9, we performed plasmid cleavage assays in the presence of one of the following divalent cations: Ca^{2+}, Mn^{2+}, Co^{2+}, Ni^{2+} and Cu^{2+}. Reactions containing Ca^{2+} yielded nicked, instead of linear plasmid (Fig 3.4A), suggesting that Ca^{2+} activates only one of the Cas9 nuclease domains. To identify which domain was activated, we assayed the single active site mutants of Cas9 (D9A or H599A) in a reaction buffer containing Ca^{2+}. We observed little cleavage with the HNH mutant (H599A) but robust cleavage with the RuvC-like mutant (D9A) (Fig 3.4B), suggesting that the HNH but not the RuvC-like domain was activated by Ca^{2+}. None of the other divalent cations tested activated either nuclease domain of Cas9 (Fig 3.4A).

Both the PAM sequence and the linker length are important for target binding

Previous studies indicate that mutations within the PAM impair DNA cleavage by Cas9 due to weakened binding (37,38,48). To determine the effect of PAM sequence and linker length on binding of LMG18311 Cas9 to DNA targets, we determined the binding affinity (K_d) of the Cas9-sgRNA complex to 5'-end labeled dsDNA targets using native gel electrophoresis (Fig 3.5A). Binding experiments were conducted with the nuclease deficient mutant of Cas9 (D9A, H599A) in the presence of Mg^{2+}. Fixed concentrations of the dsDNA targets were incubated with increasing concentrations of the Cas9-sgRNA complex (Fig 3.5A). A target containing a complementary protospacer, a 2 bp linker and a functional PAM bound to Cas9-sgRNA with an affinity of 0.94 ± 0.27 nM (Fig 3.5B).

We were unable to detect binding to a target containing a non-complementary protospacer or to a target that lacked a PAM. Mutation of the guanosine at position 1 of the PAM resulted in an ~100-fold increase in K_d (Fig 3.5B), whereas mutations at positions 2 through 5 did not significantly alter the affinity (all within ~4-fold on the consensus PAM) (Fig 3.5B). Changes in linker length had a larger effect on binding affinity (Fig 3.5C). Under the condition tested, we failed to detect binding to plasmid targets containing linker lengths of 0, 4 or 5 bps (K_d > 1000 nM), while linkers of 1 and 3 bps reduced the affinity by ~400-fold and ~20-fold, respectively (Fig 3.5C).

HNH and RuvC-like domains determine the location of their cut sites using different mechanisms

Previous studies reported Cas9 cleaves both DNA strands within the protospacer, 3 bps from its PAM proximal end, producing a predominantly blunt-end dsDNA break (7,37,38,101). To determine if linker length has any effect on where the Cas9 nuclease domains cut, we mapped the location of the cut sites in plasmids containing protospacer-1 and different lengths of linker. Following cleavage by Cas9 (programmed with an sgRNA complementary to protospacer-1), the linear plasmid products were purified by agarose gel electrophoresis and sequenced. Sequencing data revealed that the position of the cleavage site on the non-complementary strand, but not on the complementary strand, depended on linker length (Fig 3.6A). Cleavage of the complementary strand always occurred 3 nucleotides from the 5' end of the protospacer sequence, independent of the linker length (Fig 3.6A). In contrast, cleavage of the non-complementary strand occurred predominantly 5 nucleotides from the 3' end of the PAM with linker lengths of 2 or more

bps or at 4 and 5 nucleotides from the 3' end of the PAM with a linker length of 1 bp (Fig 3.6A). The site of cleavage on both strands of the DNA target was also found to be independent of spacer sequence. The location of Cas9 cut sites in plasmids containing protospacer-2 was found to be identical to plasmids containing protospacer-1 for all linker lengths investigated (Fig 3.6B). We were unable to generate enough cleaved DNA from the plasmid target with a linker length of zero for sequencing.

Discussion

Cas9, the RNA-guided endonuclease from the type II CRISPR-Cas system, has the potential to revolutionize our ability to manipulate the genomes of a wide variety of organisms (39-42,80,81,83-93,102). Targeting Cas9 to specific genomic sites relies on the presence of a PAM and complementarity between the sequence of its crRNA and the protospacer. A remarkably diverse set of PAM sequences are recognized by Cas9 orthologs (98). To date, PAM recognition and DNA cleavage have been experimentally studied in only a handful of Cas9 orthologs (37,38,96,98). Characterization of additional orthologs is expected to improve our mechanistic understanding of Cas9 and likely expand our engineering capabilities. Here we present characterization of the Cas9 protein from *S. thermophilus* LMG18311.

We demonstrate LMG18311 Cas9 is active *in vivo* through transformation assays (Fig 3.2) and *in vitro* by monitoring plasmid cleavage (Fig 3.3). We also confirm the PAM for LMG18311 Cas9 identified by sequence alignments is functional (Fig 3.1B). As observed for other Cas9 orthologs, LMG18311 Cas9 activity requires tracrRNA, crRNA and Mg^{2+} (Fig 3.3). Metal ion substitution studies also reveal Ca^{2+} likely activates the HNH but not the RuvC-like domain of LMG18311 Cas9 (Fig 3.4B). Here however we cannot rule out the possibility that the observed activation of the HNH domain may be due to trace Mg^{2+} contamination in the Ca^{2+} solution. Neither nuclease domain of *S. pyogenes* Cas9 is activated by Ca^{2+} (37).

Cas9 orthologs have been reported to cleave target DNA with a wide range of mutations in the PAM sequences (98). Yet, in natural targets PAM sequences are highly conserved. This apparent discrepancy may arise from the dual function of the PAM

(94,98). The stringency on the PAM sequence is greater for spacer acquisition than for DNA cleavage by Cas9. Consistent with this, our results show that although the PAM for LMG18311 Cas9 is conserved (Fig 3.1B), the nuclease activity of LMG18311 Cas9 tolerates a broad range of mutations in the PAM of the target DNA. Mutations to the guanosine at position 1 impair Cas9 activity, while individual mutations at positions 2 through 5 have little effect. The PAM for *N. meningitides* Cas9 also contains a single guanosine important for Cas9 activity. In addition, two recent *in vivo* studies show an AG sequence can partially replace the consensus PAM, GG, for *S. pyogenes* Cas9 (88,103). Thus, despite the varying sequence of PAM, Cas9 proteins from LMG18311, *S. pyogenes* and *N. meningitides* all contain a guanosine that appears essential for DNA silencing *in vivo*.

A previously unexplored aspect of target binding and cleavage by Cas9 is the length of the linker between the PAM and protospacer. The 41 natural targets of LMG18311 Cas9 we identified in our sequence searches all contain a 2 bp linker. However, we found that DNA containing a 3 bp linker was silenced with the same efficiency as that with a 2 bp linker (Fig 3.2D and 3F). Further lengthening or shortening of the linker eliminates CRISPR-Cas silencing and inhibits plasmid cleavage (Fig 3.2D and 3F). Thus, our results on the Type II system of *S. thermophilus* LMG18311 suggest the requirements for the length of the linker appear to be less stringent for DNA silencing than for spacer acquisition, a pattern similar to that observed for requirements on the PAM sequence.

Recognition of target DNA by either Cas9 or effector complexes from the Type I CRISPR-Cas systems is thought to be a multistep process (33,37,38,48,77). First, cellular

DNA is scanned for PAM sequences. Once a PAM is identified, the adjacent DNA duplex is destabilized enabling Cas9 to probe sequence complementarity on the target strand. Target recognition is completed if this adjacent sequence contains a protospacer that can base pair with the crRNA, stabilizing the complex. If this sequence lacks a protospacer, then the crRNA-DNA heteroduplex fails to form and Cas9 dissociates. We found the affinity of LMG18311 Cas9-sgRNA for its target sequence is ~1.0 nM (Fig 3.5), which is similar to the K_d of ~0.5 nM reported for *S. pyogenes* Cas9 (48), and comparable to the affinity of the Type I effector complexes for their DNA targets (18,74,76). Targets lacking a PAM had no detectable affinity for Cas9. As expected (37,38), the impaired nuclease activity of LMG18311 Cas9 observed when PAM sequences are mutated arises from the weakened binding affinity between Cas9 and target DNA (Fig 3.5B). Further analysis also revealed that the inhibition of cleavage of targets with different linker lengths was also due to weakened affinity (Fig 3.5C). Although both PAM and linker mutations result in reduced target affinity, they likely affect different steps in binding. PAM mutations inhibit the initial recognition of a target sequence, whereas, altering linker length likely impairs the efficiency of base pairing between crRNA and the protospacer, thus destabilizing the complex.

The length of the linker between the PAM and protospacer affects both the efficiency of DNA target cleavage and the position of the cleavage sites. This suggests the two nuclease domains of Cas9 select their cleavage sites by different mechanisms. The HNH domain cleaves the complementary strand at a fixed position while the RuvC-like domain, employing a ruler mechanism, cleaves the non-complementary strand at a

position measured from the PAM (Fig 3.6). These observations suggest the domain architecture of Cas9 is highly flexible.

Crystal structures of *S. pyogenes* and *Actinomyces naeslundii* Cas9 (43) and of *S. pyogenes* Cas9 in complex with sgRNA and its ssDNA target (44) reveal Cas9 adopts a two-lobed architecture composed of target recognition and nuclease lobes. The target recognition lobe is essential for binding the sgRNA and the complementary strand of the DNA target. The nuclease lobe contains a carboxyl-terminal domain implicated in PAM binding (43,44) as well as the HNH and RuvC-like nuclease domains. This structural organization is consistent with the two nuclease domains of Cas9 selecting their cleavage sites by different mechanisms. In the nuclease lobe, the PAM interacting and RuvC-like domains adopt a fixed position relative to each other consistent with our observation that cleavage of the non-complementary strand by the RuvC-like domain occurs at a fixed distance from the PAM (Fig 3.7). In contrast, the position of the HNH domain is highly mobile (43,44). In the current structure of Cas9-sgRNA bound to ssDNA the HNH domain is positioned away from its cleave site on the complementary strand (44). During cleavage the HNH domain must engage this strand and therefore must dock with the target recognition lobe (Fig 3.7). This docking likely determines the cleavage site of the HNH domain in the complementary strand. This is consistent with our observation that the HNH domain cleaves at a fixed position when the linker length is varied as the mobility of the HNH domain and the flexibility between the two lobes of Cas9 (43,44) likely accommodate the change in linker length (Fig 3.7).

In summary, we have characterized the substrate requirements of LMG11831 Cas9 both in vivo and in vitro. Our results enable wider target selection for genome

manipulation through the use of a distinct PAM. They also reiterate the importance of considering which Cas9 ortholog to use in genome manipulation, as those with longer PAM sequences are not necessarily more stringent in DNA cleavage. We also reveal the requirements for linker length in DNA cleavage by a Cas9 ortholog and, by varying the linker length, reveal that the two nuclease domains of Cas9 select their cut sites by different mechanisms.

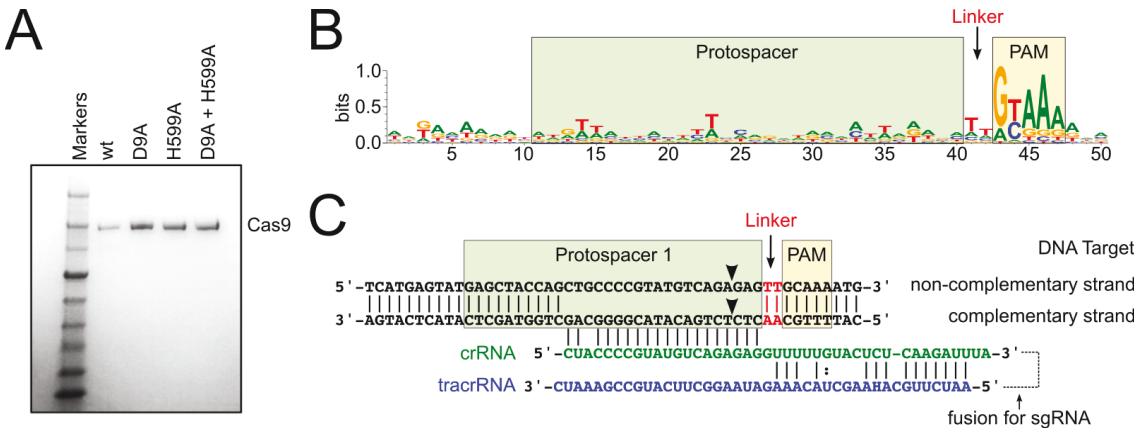

Fig 3. 1 The Type II CRISPR-Cas system of *S. thermophilus* LMG18311. (A) Coomassie stained SDS-polyacrylamide gel of Cas9, Cas9 D9A, Cas9 D599A and Cas9 D9A/D599A. (B) Logo plot revealing the PAM for LMG18311 Cas9. The position of the protospacer, PAM and linker are indicated. (C) Schematic representation of the crRNA (green), tracrRNA (blue) and DNA target (black). The position of the protospacer, PAM and linker are indicated. The site at which the crRNA and tracrRNA are fused to generate the sgRNA is indicated with a dotted line.

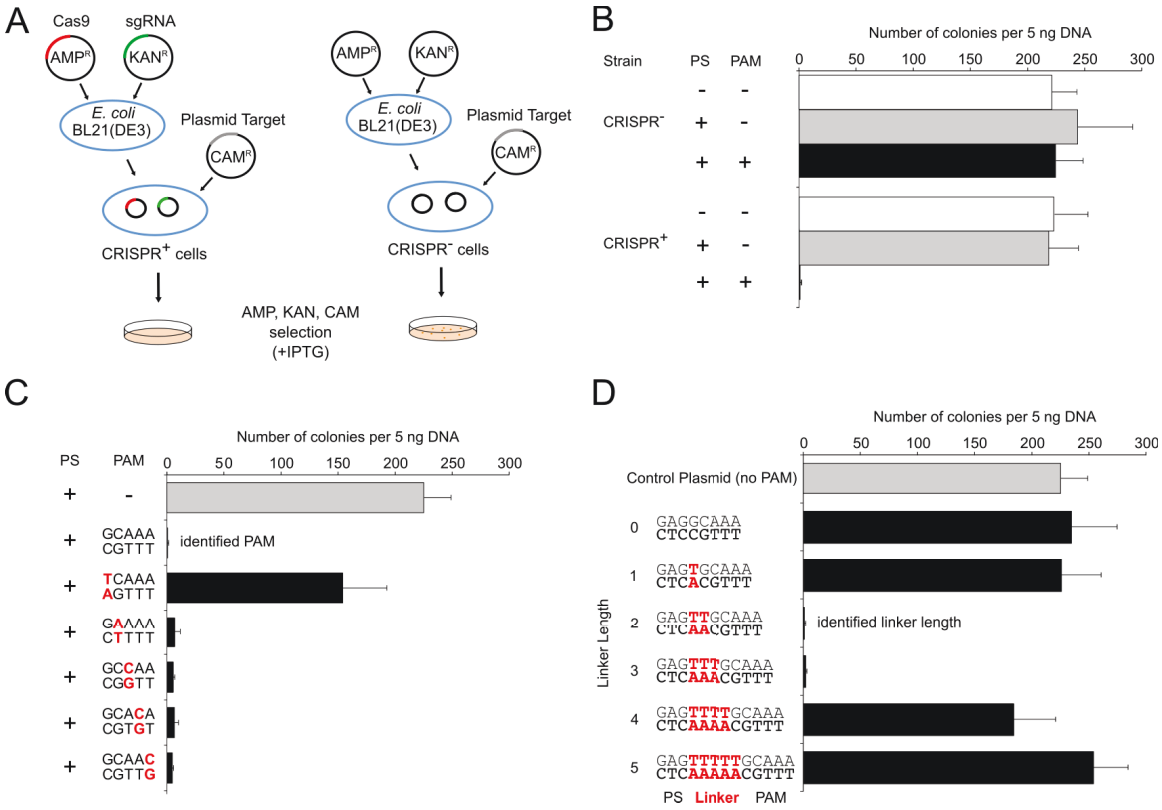

Fig 3. 2 LMG18311 Cas9 and cognate sgRNA can provide resistance to plasmid transformation in *E. coli*. (A) Schematic representation of transformation assay. (B) Interference of plasmid transformation by LMG18311 Cas9 and sgRNA in *E. coli* cells. Transformation efficiency is expressed as cfu per 5 ng of plasmid DNA. Average values from at least three biological replicates are shown, with error bars representing one standard deviation. (C) Effect of mutation in the PAM sequence on plasmid transformation efficiency. (D) Effect of linker length on plasmid transformation efficiency.

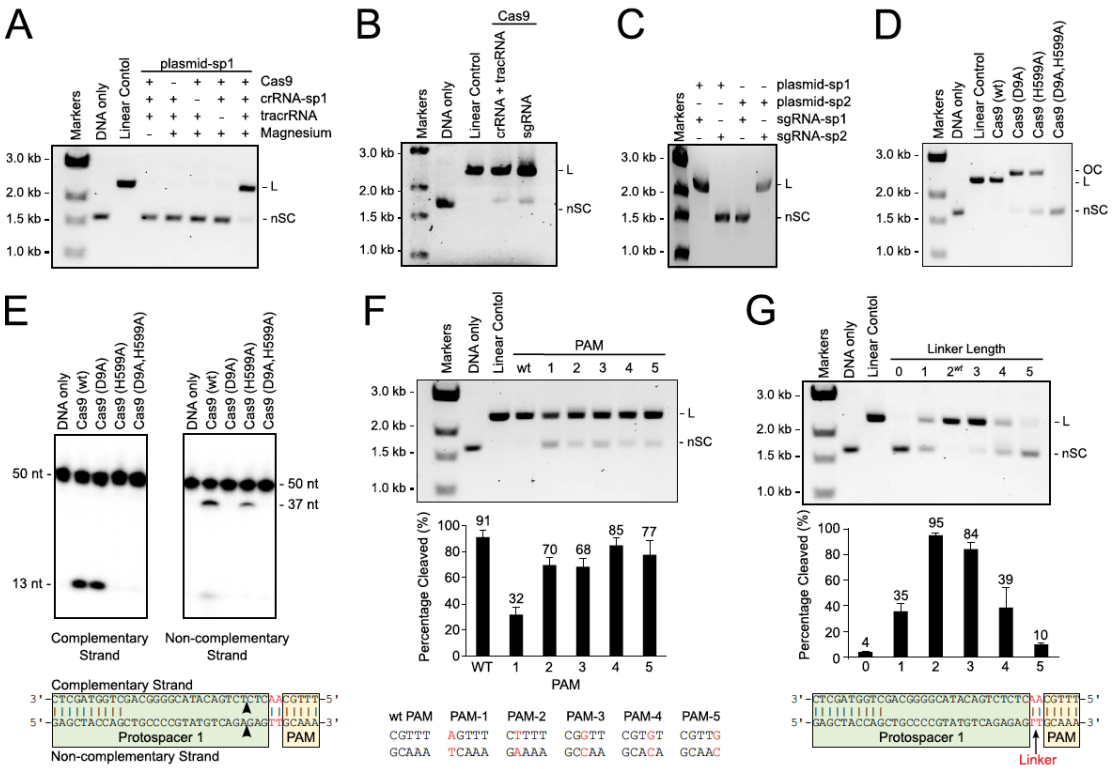

Fig 3. 3 DNA cleavage by LMG18311 Cas9 *in vitro.* (A) RNA-guided cleavage by Cas9. Reactions mixtures containing 5 nM target plasmid, 25 nM Cas9, 25 nM crRNA, 25 nM tracrRNA and 10 mM Mg^{2+} were incubated for 30 min at 37 °C. (B) A cognate sgRNA can substitute for crRNA and tracrRNA. (C) Cleavage of a plasmid target is dictated by the sgRNA sequence. (D) Cleavage of a plasmid target by active site mutants of Cas9. (E) Cleavage of a synthetic dsDNA by active site mutants of Cas9. The dsDNA was radiolabeled at the 5'-end of the complementary strand (left) or the non-complementary strand (right). Reactions were performed as in A, and products separated by 10% denaturing PAGE. The cleavage sites are indicated with arrows in the schematic diagram (bottom). (F) Cleavage of plasmid targets containing mutations in the PAM sequence. (G) Cleavage of plasmid targets containing indicated linker length. Average values from at least three biological replicates are shown, with error bars representing one standard deviation. In *A-C* and *E-F*, the position of negatively supercoiled (nSC), linear (L) and nicked or open circle (OC) plasmid is indicated. The linear control is a digestion of the plasmid target with the restriction enzyme *AgeI*.

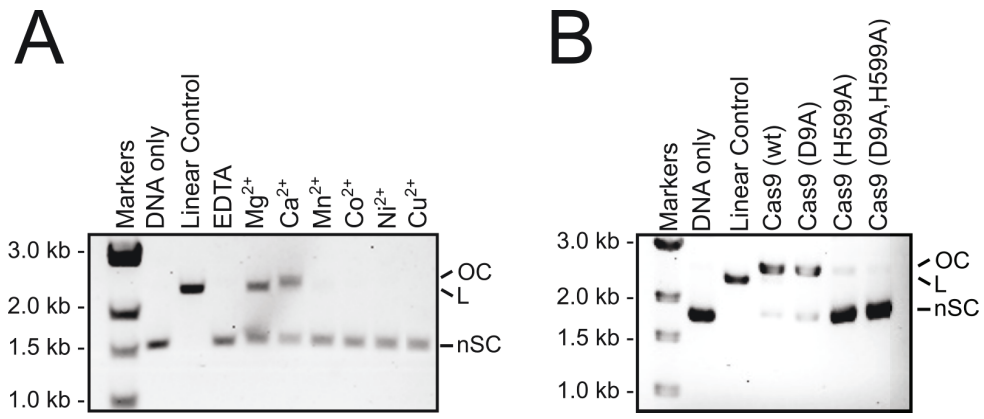

Fig 3. 4 Metal dependency of DNA cleavage by Cas9. (A) Cleavage of a target plasmid by Cas9 with either no metal or 1 mM of the indicated metal ions. All reactions were treated with 0.5 mM EDTA prior to metal addition. (B) Cleavage of a target plasmid by active site mutants of Cas9 in the presence of 10 mM Ca^{2+}. In both panels, the position of negatively supercoiled (nSC), linear (L) and nicked or open circle (OC) plasmid is indicated. The linear control is a digestion of the plasmid target with the restriction enzyme *AgeI*.

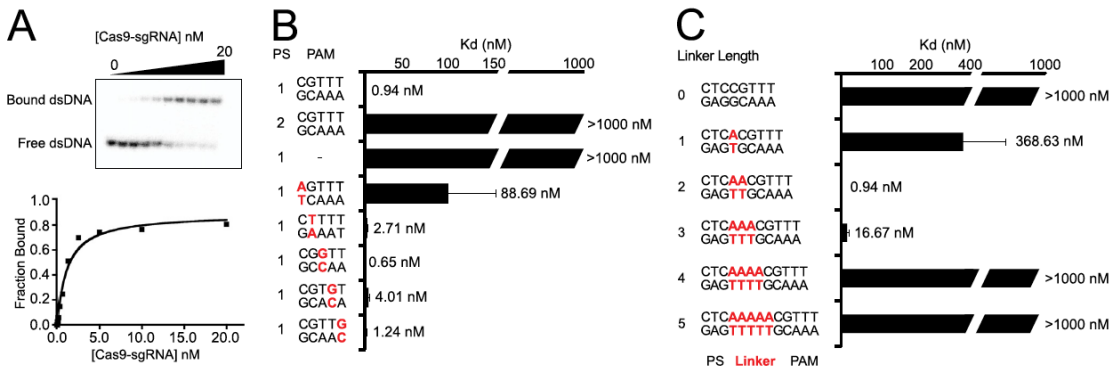

Fig 3. 5 DNA target binding by Cas9. (A) A representative gel shift assay for Cas9-sgRNA and the binding curve measured from the assay. (B, C) Bar graph plotting K_d values for (B) DNA targets with PAM mutations (labeled red) or (C) DNA targets with different linker lengths (labeled red). Average values from at least three replicates are shown, with error bars representing one standard deviation. Targets where binding was not observed are shown with K_d values at the lower limit (> 1000 nM).

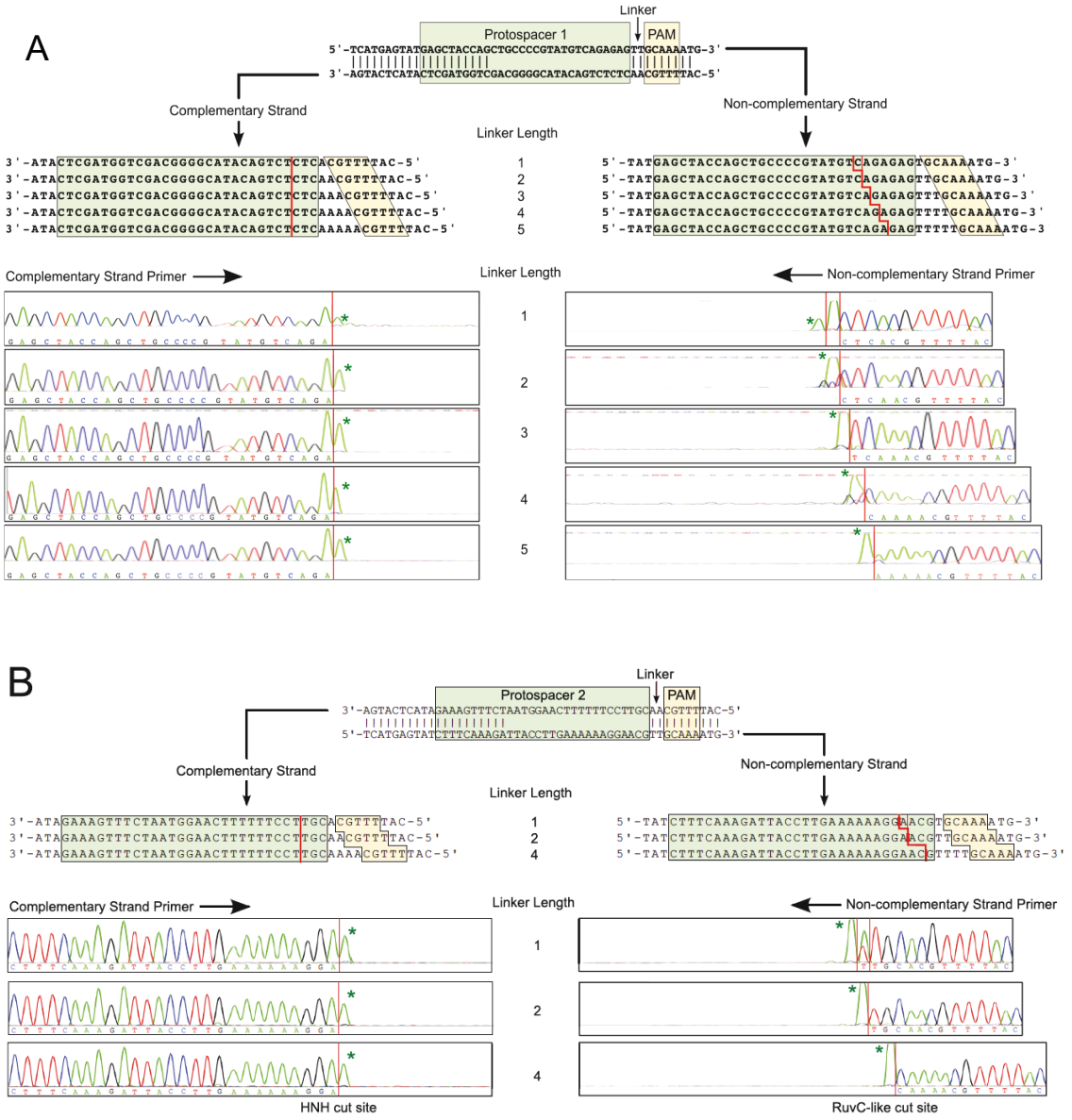

Fig 3. 6 Mapping the Cas9 cleavage sites in plasmid targets with different linker lengths. Direct sequencing electropherograms for plasmid-sp1 (A) and plasmid-sp2 (B) from complementary strand primer (bottom left) and non-complementary strand primer (bottom right) are shown. Termination of primer extension in the sequencing reaction reveals the position of the cleavage site (red line). The position of the protospacer, PAM and linker are indicated. The 3' terminal A addition, indicated by asterisk, is an artifact of the sequencing reaction.

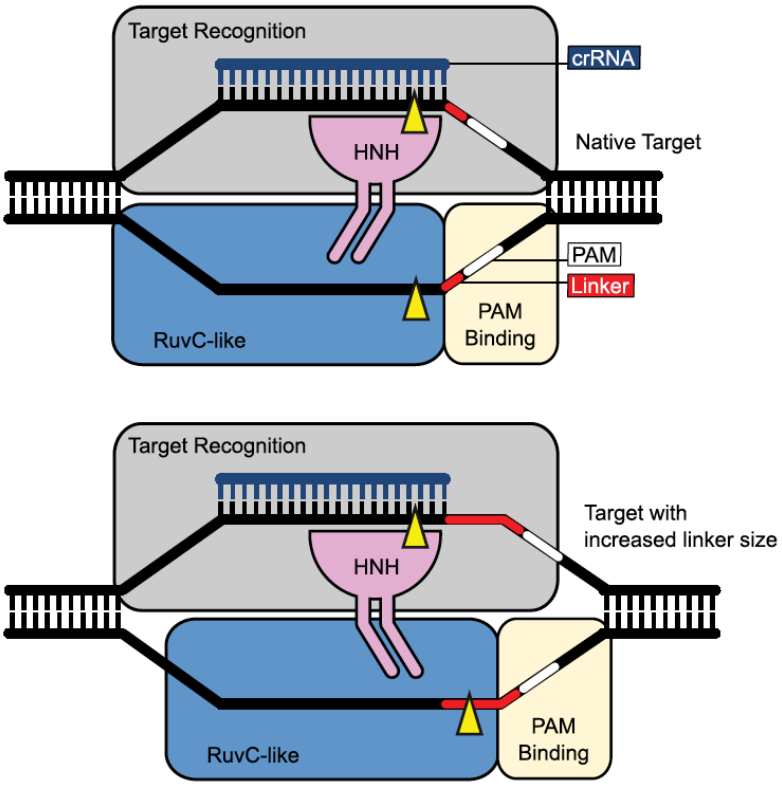

Fig 3. 7 Schematic representation of the cut site selection by HNH and RuvC-like domains of Cas9.

Materials and Methods

Identification of the PAM

Natural target sequences were found using the program BLAST. A single mismatch was allowed between the spacer and target sequences. Allowing more mismatches did not increase the number of sequences found. Sequences were considered unique if they were from distinct target genomes.

Cloning and mutagenesis

The sequence encoding full-length Cas9 was PCR-amplified from *S. thermophilus* LMG18311 genomic DNA (American Type Culture Collection) and inserted into the pMAT expression vector (27,104). The resulting construct encodes Cas9 fused to an N-terminal His_6-MBP tag. Cas9 mutants were created using QuikChange site-directed mutagenesis method (Stratagene).

To generate plasmid targets and RNA encoding vectors, synthetic oligonucleotides (100 nM), bearing the appropriate sequence, were annealed and ligated into either the pACYCDuet-1 (Novagen) or pMK-QR (GeneArt) using NcoI and EcoRI sites. Primers and oligonucleotides are listed in Table 1. All constructs were verified by DNA sequencing.

Protein expression and purification

Cas9 was overexpressed in T7Express *E. coli* (New England Biolabs). Cells were grown at 37°C in LB medium supplemented with Ampicillin to an A_{600} of ~0.3. Protein

expression was induced with 0.2 mM IPTG overnight at 20°C. Cells were harvested by centrifugation and quickly frozen in liquid nitrogen.

For purification, cells were resuspended in lysis buffer (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 10 mM imidazole, and 10% glycerol) supplemented with protease inhibitor cocktail (Sigma Aldrich) and lysed by French press. Lysate was clarified by centrifugation at 18,000 rpm at 4 °C for 45 min, and the supernatant loaded on a 5 mL immobilized metal chromatography column (Bio-Rad) charged with nickel sulfate. The column was washed with lysis buffer, and bound protein eluted with lysis buffer containing 250 mM imidazole. The elution was run on a HiLoad 26/60 S200 size exclusion column (GE Healthcare) pre-equilibrated with gel-filtration buffer A (20 mM Tris-HCl pH 8.0, and 500 mM NaCl). Fractions containing His_6-MBP tagged Cas9 were collected and treated with TEV protease overnight at 4°C to remove the His_6-MBP tag. Samples were re-applied to immobilized metal affinity chromatography resin to remove the His-tagged TEV protease, free His_6-MBP and any remaining tagged protein. The flow-through was collected, concentrated using an Ultracel 10K centrifugal filter unit (Millipore), and further purified by size exclusion chromatography in gel-filtration buffer B (20 mM Tris-HCl pH 8.0, 200 mM KCl, and 1 mM EDTA). The final fractions containing Cas9 were concentrated to ~16 mg/ml. Purified proteins were >95% pure as judged by SDS-PAGE and Coomassie staining (Fig 3.1A). The mutant variants of Cas9 were expressed and purified in the same manner as the wild-type protein (Fig 3.1A).

RNA preparation

RNAs were generated by *in vitro* transcription using T7 RNA polymerase. Plasmid templates were linearized overnight with EcoRI and then purified by phenol:chloroform extraction and ethanol precipitation. 0.5 ug of linear plasmid template was incubated with 0.1 mg/ml T7 RNA polymerase and 5 mM each of CTP, GTP, ATP, UTP in reaction buffer (25 mM Tris-HCl pH 8.0, 1.5 mM $MgCl_2$, 2 mM spermidine, 40 mM DTT) at 37 °C for 3 hours. RNA transcripts were then gel purified.

In vivo transformation assay

The recipient cells were prepared by co-transforming *E. coli* BL21 (DE3) with plasmids encoding Cas9 (pMAT) and sgRNA (pRSFDuet-1) or empty vectors. All plasmids, including the targets, had unique selection markers and origins of replication. The transformation assay was performed using the $CaCl_2$ heat-shock procedure as described in Chapter 1 with minor changes. 1 ml of cells was used for each transformation assay, and each time cells were resuspended with 0.5 ml of "Na" or "Ca solution". The final cells were resuspended in 50 µl of "Ca solution", transformed with 5 ng plasmid DNA, and recovered in LB medium containing 0.2 mM IPTG at 37 °C for 1 hour and plated on LB agar containing appropriate antibiotics and 0.2 mM IPTG. Reported transformation efficiencies are the average of at least three biological replicates. All target plasmids used in this study transformed into control recipient cells with the same efficiency (~200 colony forming units per 5 ng DNA).

Plasmid cleavage assay

Cas9 (25 nM), tracrRNA (25 nM) and crRNA (25 nM) were incubated in a cleavage buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$) at 37 °C for 30 minutes. The reactions were initiated by adding plasmid targets (4 nM), incubated at 37 °C for 30 min and quenched with phenol. The aqueous layer was extracted and separated on a 0.8% agarose gel. Gels were stained by soaking in 1× TAE buffer supplemented with 5 ug/ul ethidium bromide for 1 hour, and then for a further hour in 1× TAE buffer. Bands were visualized using an FLA-7000 (Fuji) and quantified with ImageGauge (Fuji). To account for the different binding affinity of ethidium bromide to linear and supercoiled DNA, control samples with equal amounts of DNA in both forms were loaded on the same gel. The ratios of the fluorescence intensities of linear and supercoiled bands were measured and used to calculate a correlation coefficient K (105):

$$K = \frac{I_{SC}}{I_{Lin}}$$

where I_{Lin} and I_{sc} are the intensities of the linear and supercoiled bands, respectively. In our case K was determined to be 0.4 ± 0.05 and did not vary significantly between experiments. The percentage of linear product was then calculated as follows (105):

$$Percentage\ Linear = \frac{I_{Lin}}{\frac{I_{SC}}{K} + I_{Lin}} \times 100$$

Electrophoresis mobility shift assay

DNA oligonucleotides were purified on 10% denaturing polyacrylamide gels. dsDNA targets (Table 3.1) were made by annealing each strand and purified on 12% native polyacrylamide gels containing 1× TBE. dsDNA were 5' end labeled with [γ-^{32}P]-ATP using T4 polynucleotide kinase (New England Biolabs). A fixed concentration (10-

100 pM) of labeled dsDNA targets were mixed with an increasing concentration of pre-mixed $Cas9^{D9A,H599A}$-sgRNA complex. Binding assays, performed in buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$, 0.1 mg/mL BSA, and 10% glycerol), were incubated at 37 °C for 30 min, followed by separation on 5% native polyacrylamide gels. Gels were visualized by phosphorimaging (Fuji) and quantified with ImageGauge (Fuji). Fraction DNA bound was plotted versus concentration of Cas9, and data fit to a one-site binding isotherm using GraphPad Prism software. Reported K_d values are the average of at least three replicates.

* John Choi, a ScM student (2012-2013), worked on *in vivo* transformation assay, determined correlation coefficient and generated final images for plasmid cleavage assay.

Table 3. 1 Primer and oligonucleotides used in these studies

	Sequence (5' to 3')
Primers for LMG18311 Cas9 gene amplification:	
Forward	GTGTGTCCATGGGAAGTGACTTAGTTTTAGGACTTG
Reverse	GTGTGTCTCGAGTTAAAAATCTAGCTTAGGCTTATC
Primers for site directed mutagenesis:	
D9A forward	GTGACTTAGTTTTAGGACTTGCTATCGGTATAGGTTCTGTTG
D9A Reverse	CAACAGAACCTATACCGATAGCAAGTCCTAAAACTAAGTCAC
H599A forward	CCTAATCAGTTTGAAGTAGATGCTATTTTACCTCTTTCTATCAC
H599A Reverse	GTGATAGAAAGAGGTAAAATAGCATCTACTTCAAACTGATTAGG
Oligonucleotides used to construct sgRNA plasmid:	
Forward	CATGCTACCCCGTATGTCAGAGAGGTTTTTGTACTCTGAAAAATCTT GCAGAAGCTACAAAGATAAGGCTTCATGCCGAAATC
Reverse	AATTGATTTCGGCATGAAGCCTTATCTTTGTAGCTTCTGCAAGATTT TTCAGAGTACAAAAACCTCTCTGACATACGGGGTAG
Oligonucleotides used as the template for in vitro transcription of sgRNA:	
S1 Forward	GAAATTAATACGACTCACTATAGGCTACCCCGTATGTCAGAGAGGT TTTTGTACTCTGAAAAATCTTGCAGAAGCTACAAAGATAAGGCTTC ATGCCGAAATC
S1 Reverse	GATTTCGGCATGAAGCCTTATCTTTGTAGCTTCTGCAAGATTTTTCA GAGTACAAAAACCTCTCTGACATACGGGGTAGCCTATAGTGAGTC GTATTAATTTC
S2 Forward	GAAATTAATACGACTCACTATAGGTTACCTTGAAAAAAGGAACGG TTTTTGTACTCTGAAAAATCTTGCAGAAGCTACAAAGATAAGGCTTC ATGCCGAAATC
S2 Reverse	GATTTCGGCATGAAGCCTTATCTTTGTAGCTTCTGCAAGATTTTTCA GAGTACAAAAACCGTTCCTTTTTTCAAGGTAACCTATAGTGAGTC GTATTAATTTC
RNA Oligonucleotides:	
crRNA	CUACCCCGUAUGUCAGAGAGGUUUUUGUACUCUCAAGAUUUA
tracrRNA	AAUCUUGCAGAAGCUACAAAGAUAAGGCUUCAUGCCGAAAUC
Oligonucleotides used to construct the plasmid targets:	
S1 top strand:	CATGCATTTTGCAACTCTCTGACATACGGGGCAGCTGGTAGCTCA TACTCATGA
S1 bottom strand:	AATTTCATGAGTATGAGCTACCAGCTGCCCCGTATGTCAGAGAGT TGCAAAATG
S2 top strand:	CATGCATTTTGCAACGTTCCTTTTTTCAAGGTAATCTTTGAAAGA

	TACTCATGA
S2 bottom strand:	AATTTCATGAGTATCTTTCAAAGATTACCTTGAAAAAAGGAACGT TGCAAAATG

Oligonucleotides used to construct synthetic DNA targets:

Top strand:	CATTTTGCAACTCTCTGACATACGGGGCAGCTGGTAGCTCATACTC ATGA
Bottom strand:	TCATGAGTATGAGCTACCAGCTGCCCCGTATGTCAGAGAGTTGCA AAATG

Chapter 4

Conclusions and future directions

Final Conclusions

CRISPR-Cas provides prokaryotes with adaptive and inheritable immunity against invasive genetic elements. This immunity is exerted through a three-stage pathway: adaptation, crRNA processing, and target interference. This thesis primarily focuses on the biochemical and structural aspects of the target interference stage of type I and type II systems. The studies presented here include *in vitro* and *in vivo* reconstitution of the two systems, characterization of protein-nucleic acid interactions and structural analysis of the protein complexes. These results contribute to the understanding of the molecular basis of the targeting mechanisms in type I and type II CRISPR-Cas systems.

In Chapter 2, we explored a crucial step—R-loop formation—during target recognition by type I surveillance complex Cascade. Our lab previously solved a crystal structure of *E. coli* Cascade bound to crRNA and a ssDNA target, which provided important information about target binding by Cascade (23). In this study, we identified a binding pocket for the non-target DNA strand, and verified its function through mutational analysis. We showed via filter binding assay and magnetic tweezer experiments that mutations in this pocket impair R-loop formation, probably due to the weakened interactions with the non-target strand. However, these mutations do not affect the R-loop stability once R-loops are formed, suggested by magnetic tweezer experiments and Cas3 recruitment assay. In addition, *in vivo* plasmid transformation showed that cells harboring mutant Cascade have mild defects in eliminating a target plasmid. These results suggest that, besides stabilizing the target strand through base paring with crRNA and interactions with subunits (23), sequestering the non-target strand also contributes to the formation of an R-loop. However, the R-loop stability seems to be primarily

determined by the ability of Cascade to apply a conformational "lock" (72), which is likely accompanied by a concerted structural rearrangement in Cascade upon target binding (20,23).

In Chapter 3, we characterized a Cas9 ortholog from *S. thermophilus* LMG18311. The purpose of this study was to expand current CRISPR-Cas9 genome engineering toolbox and increase the capacity of multiplex editing with different Cas9 orthologs. We demonstrated both *in vivo* and *in vitro* that LMG18311 Cas9 effectively cleaves dsDNA using either separated crRNA and tracrRNA or a chimeric sgRNA. The cleavage is dependent on sequence complementarity between crRNA and the target as well as the presence of a PAM motif. We identified the PAM as 5'-GCAAA-3', with the first G being the most important. Furthermore, we showed that both the cleavage efficiency and cleavage position vary depending on the position of PAM. The HNH domain cleaves the target strand at a fixed position, whereas the RuvC-like domain cleaves the non-target strand using a ruler mechanism. The flexibility of the relative position of the two domains is supported by the structural studies of Cas9 (43-47).

These studies, together with others, revealed a few similarities between the type I and type II systems. First of all, they both target dsDNA. Unlike type III systems, which cleave the non-template DNA strand and the RNA transcript in a transcription bubble (106-108), type I and II systems cleave dsDNA and this activity relies strictly on the DNA sequence, independent of transcription. Secondly, PAM is used in both systems to discriminate self versus "non-self" DNA elements. While in Type III systems cleavage is prevented by the presence of homology between crRNA and the 5' flanking sequence of the target (32,108), in type I and II systems the recognition of PAM adjacent to the target

licenses DNA cleavage. However, the tolerance for PAM mutations appears to be relatively relaxed. For example, *E. coli* Cascade can use CTT, CTA, CCT, CTC, or CAT for target interference (25), and *S. thermophilus* LMG18311 Cas9 tolerates mutations at position 2 to 5 of its PAM 5'-GCAAA-3'. Thirdly, both Cascade and Cas9 carry out directional formation of R-loop between the crRNA and the dsDNA target. Both systems initiate the search by scanning for PAMs. After PAM recognition, a short segment in crRNA immediately after PAM, termed "seed" sequence, is used to nucleate an R-loop, which further expands to a complete R-loop across the homologous region (37,48,76,77). Single molecule experiments have provided direct evidence for directional R-loop formation by both Cascade and Cas9 (48,72,73). In our study, we showed that the R-loop expansion in Cascade is facilitated by sequestering the non-target strand in a distinct binding pocket.

On the other hand, mechanistic differences are also observed between type I and II systems during target interference. First of all, a fundamental difference is the complex composition. Type I systems require a multimeric complex Cascade and Cas3 to complete target destruction, whereas type II systems only need a single large protein Cas9. Secondly, the RNA component in each complex differs greatly. Type I crRNA is generated by a one-step processing by Cas6 at repeat regions and thus contains an intact spacer sequence of 32-33 nt (16). In contrast, type II crRNA undergoes two processing events by RNase III and an unknown nuclease, and the resulted crRNA remains hybridized to tracrRNA at its 3' end and contains only 20 nt spacer sequence at its 5' end (36). In addition, the conformation of crRNA presented in the pre-target bound complexes is drastically different, which implies distinct mechanisms for target binding.

Interestingly, Cascade displays its entire spacer region as six discrete segments in a distorted A-form configuration (22,24), using a conformation proof-reading mechanism similar to that used by RecA (23), whereas Cas9 pre-organizes only its first 10 nt spacer sequence in A-form, using a "seed" mechanism reminiscent of that used by Argonaute proteins (46). Lastly, as for target cleavage, while type II Cas9 simply generates a blunt dsDNA break (37), type I systems involve a more complicated process. After R-loop formation by Cascade, Cas3 is recruited to the complex, first nicks the non-target strand and subsequently degrades the DNA in a 3′ to 5′ direction (27,71)

Future Directions

Cas3 recruitment

Although several structures of Cascade and Cas3 have been solved individually (see introduction), how the two interact is still unclear. Single-particle EM reveals that Cas3 colocalizes with Cse1 in Cascade, suggesting Cse1 likely provides a docking site for Cas3 (21). In fact, some systems encode a Cse1-Cas3 fusion protein, suggesting Cse1 and Cas3 functions might be closely coupled (75). But the specific interactions are yet to be defined. In addition, targets with incorrect PAMs, even when R-loops are induced, can not activate Cas3-mediated cleavage (21,73). This implies that PAM verification is a prerequisite for Cas3 activation. Structural information of Cascade (specifically Cse1) bound to correct PAMs and mutated PAMs are needed to understand the mechanism of Cas3 activation.

Priming

Cascade and Cas3 mediated interference machinery has been shown to be involved in "primed" spacer acquisition (13,14), as opposed to "naïve" acquisition where only Cas1 and Cas2 are needed. Primed spacer acquisition is driven by mutated targets that could otherwise escape the interference (13). However, the role of Cascade and Cas3 in this process is poorly understood. Single-molecule FRET experiments revealed that Cascade distinguishes mutated targets from bona fide targets using a distinct low-fidelity binding mode (109). Exactly how Cascade binds to a mutated target, for instance, containing PAM or seed mutations, has not been characterized. The canonical binding mode with bona fide targets is explained by a directional R-loop formation model. Under

this model, mismatches within PAM or seed sequence will lead to rejection of the target at an early stage. Thus, how does Cascade tolerate these escaping mutations in the primed binding mode? Furthermore, the primed binding mode does not lead to target cleavage (13). Another intriguing question is how Cascade signals Cas3 to carry out priming, but not to cleave the target. These questions require rigorous structural and biochemical studies.

Customized PAM specifications

When designing genome editing applications using Cas9, one needs to take into account the requirement of PAM adjacent to the desired target site. The most widely used *S. pyogenes* Cas9 recognizes a canonical NGG PAM. Although this sequence should appear in a genome quite frequently, it still puts constrains to applications that require high precision (110). Therefore, research endeavors have been undertaken to engineer Cas9 variants with customized PAM specifications. Initial attempts to alter PAM recognition in *S. pyogenes* Cas9 have reached some encouraging results (47,110). In a recent study, Kleinstiver *et al*. successfully generated two *S. pyogenes* Cas9 variants that recognize an NGA PAM, and showed that these variants exhibit robust editing of endogenous sites in zebrafish and human cells that are not targetable by wild-type *S. pyogenes* Cas9 (110). However, three point mutations were introduced in each case to obtain the phenotype, suggesting these alterations can be difficult to achieve. Thus, high-throughput approaches are needed to screen for mutations that can lead to desired PAM specifications. Furthermore, owing to high sequence variation in the PAM recognition

domain (45), applying similar strategies to other Cas9 orthologs will require detailed structural characterization of the particular Cas9.

Off-target effects

One major concern regarding the use of Cas9 for genome engineering applications has been its off-target effects, which means that RNA guided Cas9 can induce mutations at sites that share sequence homology to the intended target site. In the case of *S. pyogens* Cas9, the guide RNA harbors a 20-nt guide sequence, and Cas9 recognizes 2 additional nucleotides from PAM (NGG); so, if binding to a target is strictly based on the sequence, the length of 22-nt should match to unique targets in eukaryotic genomes, given that the probability is $4^{-22} \times 2$ (both DNA strands). However, it has been shown that Cas9 can tolerate up to five mismatches in the target sequence (111), greatly increasing its probability of off-target activities. Several approaches to improve the specificity of Cas9 mediated targeting have been described. One study showed that the use of synthetic guide RNA rather than guide RNA-encoding plasmids, separate crRNA and tracrRNA instead of a chimeric sgRNA, and addition of two guanine nucleotides to the 5' end of the guide sequence effectively reduces off-target mutation frequencies (112). Another study demonstrated that truncating the 5' end of guide sequence to 17 or 18 nt decreases unwanted off-target mutations by 5,000-fold while maintaining same on-target editing efficiency (113). In addition to improved RNA design, a Cas9 paired nickase strategy has been proposed, in which two Cas9 nickases (with one nuclease domain inactivated) are programmed to target adjacent sites and nick the opposite DNA strand instead of cutting both strands at one site. Although these approaches have made

great improvement in the targeting specificity, the underlying mechanism of Cas9 off-target activities has not yet been understood. A better characterization of Cas9 mediated DNA targeting is required before this technology is expanded for broader use in research or even therapeutic applications.

Appendix

Crystallization of *E. coli* Cascade bound to DNA targets

Crystallization of ssDNA bound Cascade

* This structure was previously solved by Sabin Mulepati (23). These crystals were used for DNA soaking experiments and as diffraction controls.

Cascade proteins were purified using protocol detailed in Chapter 2 Methods. ssDNA bound crystals were prepared as described in Supplementary Materials in (23). Firstly, the dsDNA target was prepared by slow-annealing 230 μM of the target strand (5'-AATCAGACAGCCCACATGGCATTCCACTTATCACTGGCAT-3') with 200 μM of a non-target strand (5'-GCCATGTGGGCTGTCTGATT-3') in a buffer containing 20 mM Tris-HCl pH 7.5, 100 mM NaCl and 0.5 mM of EDTA. The non-target strand was designed to be complementary to the 5'-region of the target strand (Fig A.1A). The dsDNA target (30 μM) was then incubated with 20 μM Cascade at 37 °C for 30 min in a buffer containing 20 mM Tris-HCl pH 8.0 and 200 mM NaCl before crystallization.

ssDNA bound Cascade crystals were obtained with sitting-drop vapor diffusion method by mixing 2 μl of the dsDNA-Cascade mixture with 1 μl of a reservoir solution containing 0.08 M sodium cacodylate pH 5.0, 0.1 M calcium acetate, and 9-11% PEG 8,000. Crystals appeared after 1 day and grew to full size within a week at 20 °C (Fig A.2A). Larger crystals (~ 500 μm x 300 μm x 300 μm) were obtained by micro-seeding.

Soaking with non-target strand DNA

Since the ssDNA-bound Cascade structure does not contain the non-target strand, we sought to soak the ssDNA-bound Cascade crystals with high concentrations of non-

target strand DNA. Because the binding to non-target strand is not sequence specific, we designed DNA substrates consisting of poly(A) and poly(T) with various lengths (dTn (n= 6, 9, 12) and dAn (n= 6, 9, 12)).

We followed a "low-salt" soaking method as published in (114) with minor modifications. Firstly, we grew ssDNA-bound Cascade crystals as described above. After a week, large crystals (≥ 250 μm x 150 μm x 150 μm) were transferred into a clean reservoir solution and gradually buffer-exchanged into a low salt solution consisting of 20 mM HEPE pH 7.5, 30 mM sodium cacodylate pH 5.0, 10% PEG 8000, and 5 % each of glycerol, sucrose, PEG 400, and ethylene glycol. Stabilized crystals were then looped into a second drop of the same low salt solution containing 5 mM ssDNA. Crystals were left in this solution overnight to allow time for binding. Crystals were then looped and flash frozen in liquid nitrogen for data collection.

Crystallization of Cascade bound to PAM-containing DNA

We attempted to introduce a double-stranded PAM into the current Cascade crystal form by extending the 3' end of the target strand such that the extended sequence could base pair with 5'-CAT-3' to form a double-stranded PAM (Fig A.1B). We designed new target strands that could fold at the 3' region as a stem-loop of various lengths (n=4,5,6 bps) (Fig A.1B). We annealed these new DNAs with the same non-target strand and carried out crystallization identically as before. Only DNA with a 5 bp stem-loop gave crystals in the initial screen. We then proceeded with this DNA and conducted further optimization. However, despite various attempts at optimization

(altering buffer concentration, pH, drop ratio, etc.), the maximum crystals obtained were ~ 200 μm x 100 μm x 100 μm (Fig A.2B). Micro-seeding did not improve the crystallization, either.

Data collection

Cascade crystals were gradually cryo-protected in 5% steps into a cryoprotectant (0.1 M sodium cacodylate pH 5.0, 0.1 M calcium acetate, 10 % PEG 8,000, and 5 % each of glycerol, sucrose, PEG 400, and ethylene glycol), and subsequently flash frozen in liquid nitrogen. For crystals soaked with ssDNA, the procedure is detailed above. Diffraction data were collected at Beamline 7-1, 11-1, and 12-2 of Stanford Synchrotron Radiation Lightsource.

Results

We had limited success with the attempts to soak in non-target DNA and to crystalize Cascade with a PAM-containing DNA. A Cascade crystal soaked with dT_{12} diffracted to ~ 5 Å. However, fitting the ssDNA-bound Cascade into density revealed no extra density of another ssDNA, suggesting no binding of the provided non-target strand. Meanwhile, the crystals of Cascade bound to PAM-containing DNA only diffracted to ~ 8 Å.

Discussion

For DNA soaking experiments, the reasons for the lack of success could be a) the binding of the non-target strand alone has very low affinity, and b) the crystallization

condition is not optimal for binding. Beloglazova *et al*. performed permanganate footprinting on dsDNA bound to *in vitro* packaged Cascade-crRNA complexes and showed no protection on the non-target strand upon Cascade binding (115). Although contradicting previous results (19,33), this result suggests that binding to the non-target strand can be transient, which could explain no detection of binding at even millimolar concentrations of DNA. The crystallization buffer contained 10% PEG8000, which is a common precipitant of DNA. In fact, precipitation of DNA was seen during preparation, and this suggests that actual DNA concentration was lower than expected.

For crystallization of Cascade bound to PAM-containing DNA, we were unable to improve the diffraction of the obtained crystals at the time. The fact that the crystals had limited growth could be due to low stability of Cascade and the DNA containing a short stem-loop. van Erp *et al*. recently showed that at least 12 additional bps on the 3' side of the protospacer sequence are required for high affinity binding to Cascade (78). Thus, the short base-paired region may cause low affinity binding to Cascade. Additionally, the 3' extension of the target DNA is positioned between Cas7.5 and Cas7.6 (78). Therefore, the bulge of the stem-loop at this region may further destabilize the binding.

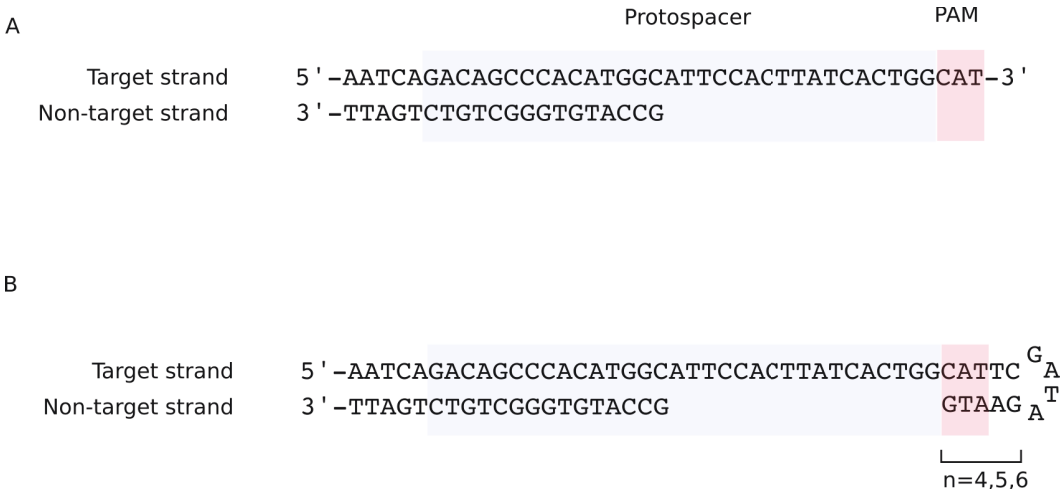

Fig A. 1 Schematic of DNA substrates used in the crystallization of Cascade bound to DNA targets. (A) DNA substrates used for crystallization of Cascade bound to ssDNA. (B) DNA substrates used for crystallization of Cascade bound to PAM-containing DNA.

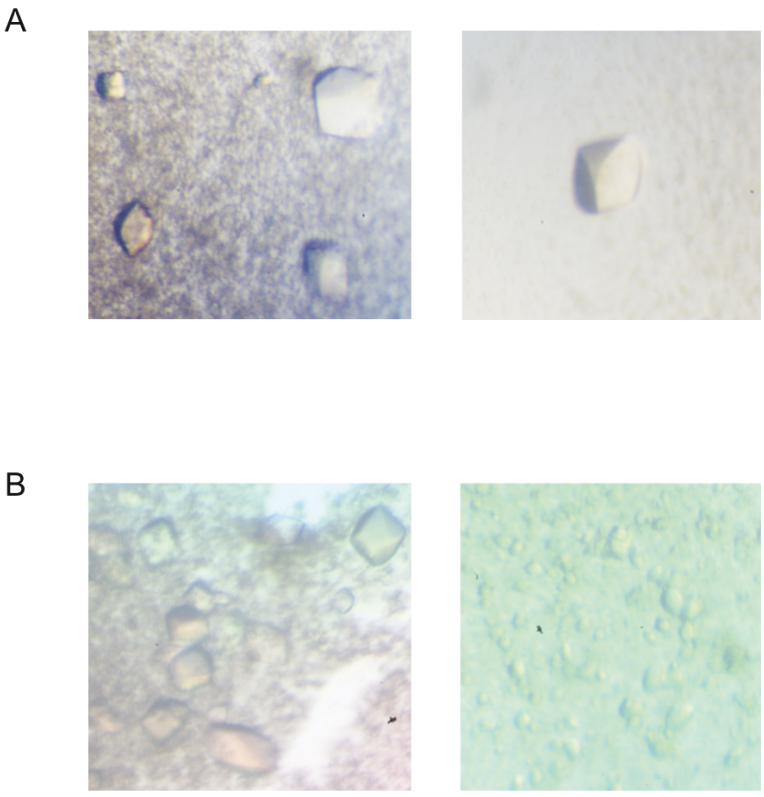

Fig A. 2 Crystals of *E. coli* Cascade bound to DNA targets. (A) Crystals of Cascade bound to ssDNA. (B) Crystals of Cascade bound to PAM-containing DNA. (Photo credit: Melesse Nune)

Crystallization of *S. thermophilus* LMG18311 Cas9

Crystallization of native crystal

S. thermophilus LMG18311 Cas9 was purified (see Chapter 3. Methods) and concentrated to ~ 16 mg/ml in gel filtration buffer (20 mM Tris-HCl pH 8.0, 200 mM KCl, and 1 mM EDTA). Crystals of Cas9 were grown with hanging drop vapor diffusion method by mixing 1 µl of protein with 1 µl of a reservoir solution (H_2O) at 20 °C. Crystals appear after ~ 2 days and grow to full size within a week. Crystal sizes were improved to ~ 800 µm × 300 µm × 200 µm with micro-seeding (Fig A.3A).

Selenomethionine (SeMet) substituted protein purification and crystallization

pMAT11 Cas9 construct was expressed in T7 Express Crystal *E. coli* cells (New England Biolabs), which are methionine auxotroph. Cells were grown overnight at 37 °C in ½ LB medium and ½ M9 minimum medium supplemented with 0.2 mg/ml ampicillin. The minimum medium was supplemented with 5 mg/L tryptophan and tyrosine, 50 mg/L each remaining amino acid other than methionine, and 1% (v/v) Kao and Michayluk Vitamin Solution (Sigma-Aldrich). The concentrated tryptophan/tyrosine stock was made in 200 mM HCl, whereas the other amino acids were dissolved in water. The next day, a 10 mL overnight culture was used to inoculate 1 L of the same minimal media without LB supplemented with 50 mg/ml SeMet and 0.2 mg/ml ampicillin. 0.2 mM IPTG was added to induce protein expression at the beginning of the growth. Cells were grown at 37 °C to an OD_{600} of 0.4~0.5, and continued to grow for 18 h at 20 °C. Cells were harvested by centrifugation and quickly frozen in liquid nitrogen. SeMet Cas9 was purified similarly as native Cas9 but with minor modifications. All buffers for SeMet

Cas9 purification contained 1 mM TCEP. In addition, lysis buffer contained 1M NaCl instead of 500 mM NaCl for a better removal of nucleic acid contamination.

SeMet Cas9 crystals were grown in the identical crystallization condition as the native crystals. However, crystals grew to a limited size (~ 300 μm × 150 μm × 100 μm) with morphology different from native crystals (Fig A.3B). Attempts to improve the crystals by adding crystal seeds, salt ($MgCl_2$, KCl), or reducing agent (β-ME, DTT, TCEP, and Tris(hydroxymethyl)phosphine) did not achieve much success.

Heavy metal soaking

Native crystals were first harvested in cryoprotectant containing 2 mM Tris-HCl pH 8.0, 10 mM KCl, and 35% glycerol. The crystals were then soaked in cryoprotectant supplemented with 100 μM ~ 1 mM heavy-metal compounds for various duration of time (Table A.1). Additionally, because mercury binding to cysteine residues may cause significant nonisomorphism (116), we generated two Cas9 mutants C329S and C1086S. Crystals of mutants were soaked primarily with mercury compounds for ~ 3 h (Table A.1).

Data collection

Crystals were gradually buffer-exchanged into a cryoprotectant containing 2 mM Tris-HCl pH 8.0, 10 mM KCl, and 35% glycerol, and flash frozen in liquid nitrogen. Diffraction data were collected at Beamline 7-1, 11-1, and 12-2 of Stanford Synchrotron Radiation Lightsource.

Results

Native crystals diffracted to 3.3 Å; SeMet derivative crystals diffracted to ~7 Å. Datasets obtained with heavy metal soaked crystals showed no detectable anomalous signal. The crystals were identified as belonging to space group I222, with one molecule per asymmetric unit. One unit cell is 82 × 174 × 270 $Å^3$. However, the heavy atom derivatives were unable to provide useful phasing information needed to solve the structure.

Discussion

We encountered several technical problems with this project. Firstly, Cas9 crystals grew in low ionic conditions such as H_2O. As a result, the crystals were sensitive to any additives such as buffering agent or salt, which largely limited the strategies to improve crystal growth. Secondly, the crystals diffracted poorly and the diffraction was highly anisotropic. Although we were able to improve crystal size and morphology, diffraction quality did not improve accordingly. The best dataset obtained so far was at 3.3 Å. Thirdly, SeMet protein purification gave very poor yields, and the SeMet crystals were not easily reproducible. Fourthly, because of the high sensitivity, native crystals could only sustain submillimolar concentrations of heavy metal solution, which could explain the ineffective binding. But in general, crystals were more sensitive to inorganic compounds such as $HgCl_2$ or $(NH_4)_2HgCl_2$, and less sensitive to organic compounds such as thiomersal ($C_9H_9HgNaO_2S$) or C_2H_5HgCl. Lastly, molecular replacement with two known Cas9 structures (4CMP and 4OGE) (43) at the time did not achieve much success due to low sequence identity. These published structures also lowered the priority of this

project.

To determine this structure, several strategies could be tried in the future: a) seek a different crystal form; b) further optimize SeMet protein prep (growth, expression, buffer, etc.) and SeMet crystals; c) perform more exhaustive molecular replacement searches to currently available Cas9 structures (4CMP, 4OGE, 4ZT0, 4UN3, 5CZZ).

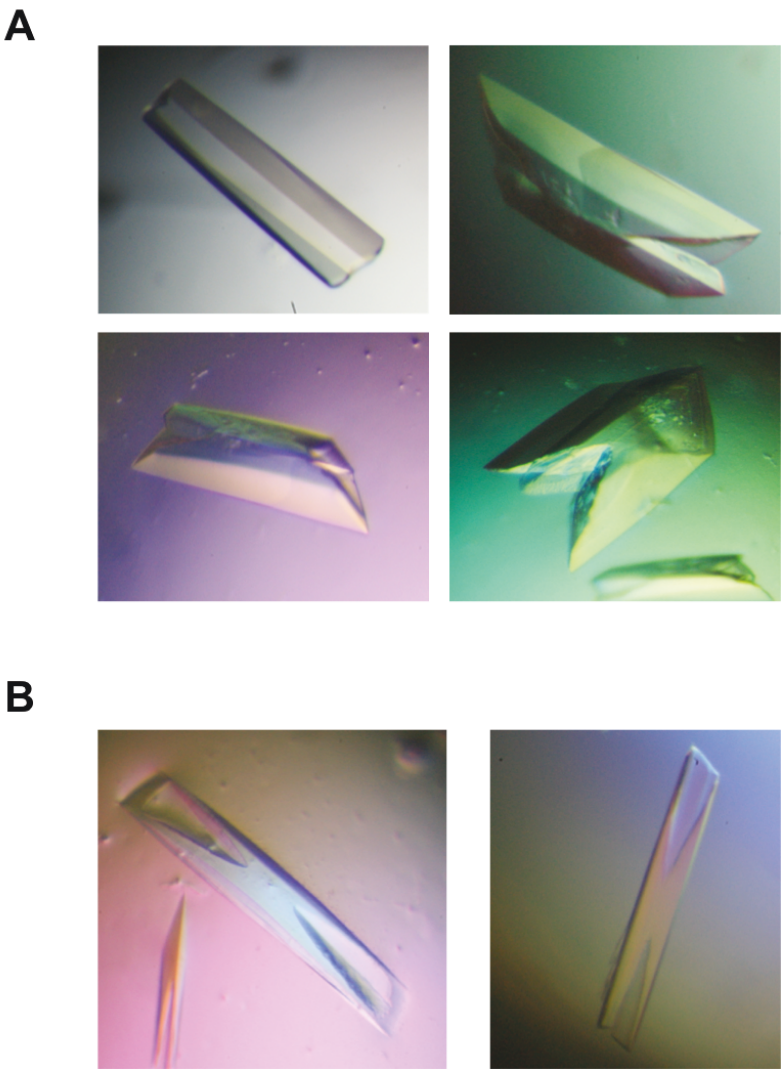

Fig A. 3 Crystals of *S. thermophilus* LMG18311 Cas9. (A) Examples of native Cas9 crystals. (B) Examples of SeMet derivative Cas9 crystals.

Table A. 1 Heavy metal compounds used for Cas9 crystal soaks

Number	Metal	Compound
1	Hg	Mercury chloride
2	Hg	Mercury acetate
3	Hg	Mercury (III) potassium iodide
4	Hg	Mercury ammonium chloride
5	Hg	Ethyl mercuric chloride
6	Hg	4-(chloromercuri) benzene-sulfonic acid
7	Hg	4-chloromercuribenzoic acid
8	Hg	Thimerosal
9	Pt	cis-Platinum(II) diammine dichloride
10	Pt	Potassium hexabromo platinate (IV)
11	W	Sodium tungstate
12	W	Paratungstate cluster
13	Ta	Tantalum cluster
14	Sm	Samarium (III) chloride
15	Eu	Europium (III) chloride
16	Yb	Ytterbium (III) chloride
17	Ho	Holmium (III) chloride
18	Ho	Holmium (III) acetate
19	Er	Erbium(III) chloride
20	I	“Magic Triangle” I3C

References

1. Bergh O, Børsheim KY, Bratbak G, Heldal M. High abundance of viruses found in aquatic environments. Nature. Nature Publishing Group; 1989 Aug 10;340(6233):467–8.

2. Samson JE, Magadán AH, Sabri M, Moineau S. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol. 2013 Oct;11(10):675–87.

3. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol. Springer-Verlag; 2005 Feb;60(2):174–82.

4. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology (Reading, Engl). Microbiology Society; 2005 Aug;151(Pt 8):2551–61.

5. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology (Reading, Engl). Microbiology Society; 2005 Mar;151(Pt 3):653–63.

6. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. American Association for the Advancement of Science; 2007 Mar 23;315(5819):1709–12.

7. Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010 Nov 4;468(7320):67–71.

8. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol. Nature Publishing Group; 2015 Sep 28.

9. van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat Rev Microbiol. Nature Publishing Group; 2014 Jul;12(7):479–92.

10. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011 Jun;9(6):467–77.

11. Nuñez JK, Lee ASY, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature. 2015 Mar

12;519(7542):193–8.

12. Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, et al. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell. Elsevier Inc; 2015 Oct 14;:1–15.

13. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Comms. 2012 Jul 10;3:945.

14. Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ. CRISPR interference directs strand specific spacer acquisition. PLoS ONE. Public Library of Science; 2012;7(4):e35888.

15. Fineran PC, Charpentier E. Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. Virology. 2012 Dec 20;434(2):202–9.

16. Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. American Association for the Advancement of Science; 2008 Aug 15;321(5891):960–4.

17. Gesner EM, Schellenberg MJ, Garside EL, George MM, MacMillan AM. Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nature Publishing Group. Nature Publishing Group; 2011 May 15;18(6):688–92.

18. Sashital DG, Jinek M, Doudna JA. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nature Publishing Group. Nature Publishing Group; 2011 May 15;18(6):680–7.

19. Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, et al. Structural basis for CRISPR RNA-guided DNA recognition by Cascade 2011. Nature Publishing Group. Nature Publishing Group; 2011 Apr 3;18(5):529–36.

20. Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJJ, van der Oost J, et al. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature. Nature Publishing Group; 2012 Apr 12;477(7365):486–9.

21. Hochstrasser ML, Taylor DW, Bhat P, Guegler CK, Sternberg SH, Nogales E, et al. CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc Natl Acad Sci USA. National Acad Sciences; 2014 May 6;111(18):6618–23.

22. Jackson RN, Golden SM, van Erp PBG, Carter J, Westra ER, Brouns SJJ, et al. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science. American Association for the

Advancement of Science; 2014 Sep 19;345(6203):1473–9.

23. Mulepati S, Héroux A, Bailey S. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science. American Association for the Advancement of Science; 2014 Sep 19;345(6203):1479–84.

24. Zhao H, Sheng G, Wang J, Wang M, Bunkoczi G, Gong W, et al. Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature. Nature Publishing Group; 2014 Nov 6;515(7525):147–50.

25. Fineran PC, Gerritzen MJH, Suárez-Diez M, Künne T, Boekhorst J, van Hijum SAFT, et al. Degenerate target sites mediate rapid primed CRISPR adaptation. Proc Natl Acad Sci USA. National Acad Sciences; 2014 Apr 22;111(16):E1629–38.

26. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO Journal. Nature Publishing Group; 2011 Feb 22;30(7):1335–42.

27. Mulepati S, Bailey S. In Vitro Reconstitution of an Escherichia coli RNA-guided Immune System Reveals Unidirectional, ATP-dependent Degradation of DNA Target. Journal of Biological Chemistry. 2013 Aug 2;288(31):22184–92.

28. Mulepati S, Bailey S. Structural and Biochemical Analysis of Nuclease Domain of Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated Protein 3 (Cas3). Journal of Biological Chemistry. 2011 Sep 2;286(36):31896–903.

29. Gong B, Shin M, Sun J, Jung C-H, Bolt EL, van der Oost J, et al. Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. Proc Natl Acad Sci USA. National Acad Sciences; 2014 Nov 18;111(46):16359–64.

30. Huo Y, Nam KH, Ding F, Lee H, Wu L, Xiao Y, et al. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nature Publishing Group. Nature Publishing Group; 2014 Sep;21(9):771–7.

31. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology (Reading, Engl). 2009 Mar;155(Pt 3):733–40.

32. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature. 2010 Jan 28;463(7280):568–71.

33. Sashital DG, Wiedenheft B, Doudna JA. Mechanism of foreign DNA selection in a bacterial adaptive immune system. Molecular Cell. 2012 Jun 8;46(5):606–15.

34. Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, et al. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature. 2015 Mar 12;519(7542):199–202.

35. Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev. Cold Spring Harbor Lab; 2015 Feb 15;29(4):356–61.

36. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. 2011 Mar 31;471(7340):602–7. Available from: http://www.nature.com/doifinder/10.1038/nature09886

37. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. 2012 Aug 16;337(6096):816–21. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1225829

38. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci USA. National Acad Sciences; 2012 Sep 25;109(39):E2579–86.

39. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science. 2013 Feb 14;339(6121):819–23.

40. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. American Association for the Advancement of Science; 2013 Feb 15;339(6121):823–6.

41. Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. Elife. 2013;2:e00471.

42. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-SpecificControl of Gene Expression. Cell. Elsevier; 2013 Feb 28;152(5):1173–83.

43. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science. 2014 Mar 14;343(6176):1247997–7.

44. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, et al. Crystal Structure of Cas9in Complex with Guide RNA and Target DNA. Cell. Elsevier Inc; 2014 Feb 27;156(5):935–49.

45. Nishimasu H, Cong L, Yan WX, Ran FA, Zetsche B, Li Y, et al. Crystal Structure of Staphylococcus aureus Cas9. Cell. 2015 Aug 27;162(5):1113–26.

46. Jiang F, Zhou K, Ma L, Gressel S, Doudna JA. A Cas9–guide RNA complex preorganized for target DNA recognition. Science. American Association for the Advancement of Science; 2015 Jun 26;348(6242):1477–81.

47. Anders C, Niewoehner O, Duerst A, Jinek M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature. Nature Publishing Group; 2014 Sep 16;513(7519):569–73.

48. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014 Mar 6;507(7490):62–7.

49. Pennisi E. The CRISPR craze. Science. American Association for the Advancement of Science; 2013 Aug 23;:833–6.

50. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014 Jun 5;157(6):1262–78.

51. Charpentier E, Marraffini LA. Harnessing CRISPR-Cas9 immunity for genetic engineering. Curr Opin Microbiol. 2014 Jun;19:114–9.

52. Shariat N, Dudley EG. CRISPRs: molecular signatures used for pathogen subtyping. Appl Environ Microbiol. American Society for Microbiology; 2014 Jan;80(2):430–9.

53. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, et al. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. Journal of Bacteriology. 2008 Jan 29;190(4):1401–12.

54. Fabre L, Zhang J, Guigon G, Le Hello S, Guibert V, Accou-Demartin M, et al. CRISPR typing and subtyping for improved laboratory surveillance of Salmonella infections. Mokrousov I, editor. PLoS ONE. Public Library of Science; 2012;7(5):e36995.

55. Liu F, Barrangou R, Gerner-Smidt P, Ribot EM, Knabel SJ, Dudley EG. Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of Salmonella enterica subsp. enterica. Appl Environ Microbiol. American Society for Microbiology; 2011 Mar;77(6):1946–56.

56. Shariat N, DiMarzio MJ, Yin S, Dettinger L, Sandt CH, Lute JR, et al. The combination of CRISPR-MVLST and PFGE provides increased discriminatory power for differentiating human clinical isolates of Salmonella enterica subsp. enterica serovar Enteritidis. Food Microbiol. 2013 May;34(1):164–73.

57. Bikard D, Hatoum-Aslan A, Mucida D, Marraffini LA. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. Cell Host Microbe. 2012 Aug 16;12(2):177–86.

58. Hatoum-Aslan A, Marraffini LA. Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens. Curr Opin Microbiol. 2014 Feb;17:82–90.

59. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. American Association for the Advancement of Science; 2008 Dec 19;322(5909):1843–5.

60. Palmer KL, Gilmore MS. Multidrug-resistant enterococci lack CRISPR-cas. MBio. American Society for Microbiology; 2010;1(4):e00227–10.

61. Marraffini LA. CRISPR-Cas immunity against phages: its effects on the evolution and survival of bacterial pathogens. Heitman J, editor. PLoS Pathog. Public Library of Science; 2013;9(12):e1003765.

62. Luo ML, Leenay RT, Beisel CL. Current and future prospects for CRISPR-based tools in bacteria. Biotechnol Bioeng. 2015 Oct 13.

63. Bikard D, Euler CW, Jiang W, Nussenzweig PM, Goldberg GW, Duportet X, et al. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol. 2014 Nov;32(11):1146–50.

64. Citorik RJ, Mimee M, Lu TK. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat Biotechnol. 2014 Nov;32(11):1141–5.

65. Schiffer JT, Aubert M, Weber ND, Mintzer E, Stone D, Jerome KR. Targeted DNA mutagenesis for the cure of chronic viral infections. J Virol. American Society for Microbiology; 2012 Sep;86(17):8920–36.

66. Kennedy EM, Cullen BR. Bacterial CRISPR/Cas DNA endonucleases: A revolutionary technology that could dramatically impact viral research and treatment. Virology. 2015 May;479-480:213–20.

67. Ebina H, Misawa N, Kanemura Y, Koyanagi Y. Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus. Sci Rep. 2013 Aug 26;3.

68. Hu W, Kaminski R, Yang F, Zhang Y, Cosentino L, Li F, et al. RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. Proc Natl Acad Sci USA. National Acad Sciences; 2014 Aug 5;111(31):11461–6.

69. Xiao-Jie L, Hui-Ying X, Zun-Ping K, Jin-Lian C, Li-Juan J. CRISPR-Cas9: a new and promising player in gene therapy. J Med Genet. BMJ Publishing Group Ltd; 2015 May;52(5):289–96.

70. Xue W, Chen S, Yin H, Tammela T, Papagiannakopoulos T, Joshi NS, et al. CRISPR-mediated direct mutation of cancer genes in the mouse liver. Nature. Nature Publishing Group; 2014 Oct 16;514(7522):380–4.

71. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. The EMBO Journal. 2013 Feb 6;32(3):385–94.

72. Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T, Pschera P, et al. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc Natl Acad Sci USA. National Acad Sciences; 2014 Jul 8;111(27):9798–803.

73. Rutkauskas M, Sinkunas T, Songailiene I, Tikhomirova MS, Siksnys V, Seidel R. Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. Cell Rep. 2015 Mar 3.

74. Mulepati S, Orr A, Bailey S. Crystal Structure of the Largest Subunit of a Bacterial RNA-guided Immune Complex and Its Role in DNA Target Binding. Journal of Biological Chemistry. 2012 Jun 29;287(27):22445–9.

75. Westra ER, van Erp PBG, Künne T, Wong SP, Staals RHJ, Seegers CLC, et al. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Molecular Cell. 2012 Jun 8;46(5):595–605.

76. Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proceedings of the National Academy of Sciences. National Acad Sciences; 2011 Jun 21;108(25):10092–7.

77. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, et al. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci USA. National Acad Sciences; 2011 Jun 21;108(25):10098–103.

78. van Erp PBG, Jackson RN, Carter J, Golden SM, Bailey S, Wiedenheft B. Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli. Nucleic Acids Research. 2015 Aug 3;:gkv793.

79. Ban C, Junop M, Yang W. Transformation of MutL by ATP binding and hydrolysis: a switch in DNA mismatch repair. Cell. 1999 Apr 2;97(1):85–97.

80. Shen B, Zhang J, Wu H, Wang J, Ma K, Li Z, et al. Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. Cell Res. Nature Publishing Group; 2013 May;23(5):720–3.

81. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, et al. One-

step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell. 2013 May 9;153(4):910–8.

82. Cho SW, Kim S, Kim JM, Kim J-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol. Nature Publishing Group; 2013 Mar;31(3):230–2.

83. Li J-F, Norville JE, Aach J, McCormack M, Zhang D, Bush J, et al. Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. Nat Biotechnol. Nature Publishing Group; 2013 Aug;31(8):688–91.

84. Nekrasov V, Staskawicz B, Weigel D, Jones JDG, Kamoun S. Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. Nat Biotechnol. Nature Publishing Group; 2013 Aug;31(8):691–3.

85. Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, et al. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. Genetics. Genetics Society of America; 2013 Aug;194(4):1029–35.

86. Friedland AE, Tzur YB, Esvelt KM, Colaiácovo MP, Church GM, Calarco JA. Heritable genome editing in C. elegans via a CRISPR-Cas9 system. Nat Meth. Nature Publishing Group; 2013 Aug;10(8):741–3.

87. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Research. Oxford University Press; 2013 Apr;41(7):4336–43.

88. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. Nature Publishing Group; 2013 Mar;31(3):233–9.

89. Nakayama T, Fish MB, Fisher M, Oomen Hajagos J, Thomsen GH, Grainger RM. Simple and efficient CRISPR/Cas9‐mediated targeted mutagenesis in Xenopus tropicalis. genesis. 2013 Dec 1;51(12):835–43.

90. Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol. Nature Publishing Group; 2013 Mar;31(3):227–9.

91. Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Research. Oxford University Press; 2013 Aug;41(15):7429–37.

92. Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK. CRISPR RNA-guided activation of endogenous human genes. Nat Meth. Nature Publishing

Group; 2013 Oct;10(10):977–9.

93. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013 Jul 18;154(2):442–51.

94. Shah SA, Erdmann S, Mojica FJM, Garrett RA. Protospacer recognition motifs: mixed identities and functional diversity. rnabiology. Taylor & Francis; 2013 May;10(5):891–9.

95. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. Journal of Bacteriology. 2008 Jan 29;190(4):1390–400.

96. Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, et al. Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis. Molecular Cell. Elsevier Inc; 2013 May 23;50(4):488–503.

97. Mali P, Esvelt KM, Church GM. Cas9 as a versatile tool for engineering biology. Nat Meth. 2013 Sep 27;10(10):957–63.

98. Esvelt KM, Mali P, Braff JL, Moosburner M, Yaung SJ, Church GM. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. Nat Meth. 2013 Sep 29;10(11):1116–21.

99. Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Research. 2011 Nov 23;39(21):9275–82.

100. Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V. crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. rnabiology. Taylor & Francis; 2013 May;10(5):841–51.

101. Magadán AH, Dupuis M-È, Villion M, Moineau S. Cleavage of Phage DNA by the Streptococcus thermophilus CRISPR3-Cas System. Poteete AR, editor. PLoS ONE. Public Library of Science; 2012 Jul 20;7(7):e40913.

102. Cho SW, Kim S, Kim JM, Kim J-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol. Nature Publishing Group; 2013 Jan 29;31(3):230–2.

103. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. Nature Publishing Group; 2013 Sep;31(9):827–32.

104. Peränen J, Rikkonen M, Hyvönen M, Kääriäinen L. T7 vectors with modified T7lac promoter for expression of proteins in Escherichia coli. Anal Biochem.

1996 May 1;236(2):371–3.

105. Panyutin IV, Luu AN, Panyutin IG, Neumann RD. Strand breaks in whole plasmid dna produced by the decay of (125)I in a triplex-forming oligonucleotide. Radiat Res. 2001 Aug;156(2):158–66.

106. Deng L, Garrett RA, Shah SA, Peng X, She Q. A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Molecular Microbiology. 2013 Mar;87(5):1088–99.

107. Goldberg GW, Jiang W, Bikard D, Marraffini LA. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature. 2014 Oct 30;514(7524):633–7.

108. Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell. 2015 May 21;161(5):1164–74.

109. Blosser TR, Loeff L, Westra ER, Vlot M, Künne T, Sobota M, et al. Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. Molecular Cell. 2015 Apr 2;58(1):60–70.

110. Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, Zheng Z, et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature. Nature Publishing Group; 2015 Jul 23;523(7561):481–5.

111. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. Nature Publishing Group; 2013 Sep;31(9):822–6.

112. Cho SW, Kim S, Kim Y, Kweon J, Kim HS, Bae S, et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res. Cold Spring Harbor Lab; 2014 Jan;24(1):132–41.

113. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol. 2014 Mar;32(3):279–84.

114. Duderstadt KE, Chuang K, Berger JM. DNA stretching by bacterial initiators promotes replication origin opening. Nature. Nature Publishing Group; 2011 Oct 13;478(7368):209–13.

115. Beloglazova N, Kuznedelov K, Flick R, Datsenko KA, Brown G, Popovic A, et al. CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli. Nucleic Acids Research. Oxford University Press; 2015 Jan 9;43(1):530–43.

116. Cook WJ, Jeffrey LC, Sullivan ML, Vierstra RD. Three-dimensional structure of

a ubiquitin-conjugating enzyme (E2). Journal of Biological Chemistry. 1992 Jul 25;267(21):15116–21.

Hongfan Chen

DOB Dec 27, 1987
Hangzhou, Zhejiang, China
Email: fdchenhongfan@gmail.com

EDUCATION

2010 – present	**Ph.D.** candidate, Biochemistry & Molecular Biology, expected Jan 2016 Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
2006 – 2010	**B.S.**, Biotechnology Fudan University, Shanghai, China
Summer 2009	Attendee Harvard Summer School, Shanghai, China
Fall 2008	Exchange student University of Hong Kong, Hong Kong, China

RESEARCH EXPERIENCE

2010 – present	**Research Assistant, Johns Hopkins Bloomberg School of Public Health** Department of Biochemistry & Molecular Biology, PI: Scott Bailey Project: Biochemical and structural characterization of RNA-guided DNA targeting CRISPR/Cas systems in bacteria.
2009 – 2010	**Research Assistant, Fudan University** Center for Biotechnology, PI: Chunhua Yin Project: Synthesis and characterization of chitosan-derivative nanoparticles as anti-tumor drug delivery platforms.
2008	**Undergraduate Researcher, Fudan University** Institute of Genetics, PI: Hongyan Wang Project: Evaluation of methods for human nuclear DNA extraction from free margins of finger nail and hair samples.

PUBLICATIONS

Chen, H., Rouillon, C., Kaila, J., Mallon, J., Seidel, R., Bailey, S.. Mechanistic insights into R-loop formation by the E. coli surveillance complex Cascade. In preparation

Chen, H., Choi, J., Bailey, S.. Cut site selection by the two nuclease domains of the Cas9 RNA-guided endonuclease. J Biol Chem. 2014 May 9;289(19):13284-94.

Zhang, H., **Chen, H.**, An, Y., Wang, H., Duan, W.. Evaluation on methods of human nuclear DNA extraction from free margins of finger nail and hair samples. Chinese Journal of Evidence Based Pediatrics 4(5) 431-435, 2009

PRESENTATIONS

Chen, H., Rouillon, C., Seidel, R., Bailey, S. Insights into R-loop formation by *E. coli* Cascade. Poster presentation, CRISPR Conference 2015, New York, NY

Chen, H., Bailey, S. Structure-function studies Structure-function Studies of *E. coli* RNA-guided surveillance complex Cascade. Oral presentation, JHSPH BMB Annual Retreat 2015, Baltimore, MD

Chen, H., Choi, J., Bailey, S. Cleavage site selection by Cas9 nuclease domains. Poster presentation, JHSPH BMB Annual Retreat 2014, Baltimore, MD

Chen, H., Bailey, S. Biochemical Characterization of *Streptococcus thermophilus* Cas9. Oral presentation, JHSPH BMB Colloquium 2013, Baltimore, MD

Chen, H., Choi, J., Bailey, S. Functional study of Cas9 in RNA-guided DNA cleavage. Poster presentation, JHSPH BMB Annual Retreat 2013, Baltimore, MD

Chen, H., Bailey, S. Functional studies of Type II CRISPR/Cas system. Oral presentation, JHSPH BMB Annual Retreat 2012, Baltimore, MD

Chen, H., Estrella, M., Mulepati, S., Bailey, S. CRISPR: the prokaryotic defense system. JHU Institute for Biophysical Research Annual Retreat 2011, Baltimore, MD

MENTORING & TEACHING EXPERIENCES

Fall, 2015	Elmer A. Zapata-Mercado (Biophysics Rotation student, co-mentored with John Mallon)
Spring, 2015	Melesse Nune (Biophysics Rotation student)
Fall, 2014	Miranda Russo (Biophysics Rotation student)
Fall, 2013	Jessica Hopkins (BMB Rotation student)
Fall, 2011	Tutor, *Biophysical and Biochemical Principles*

SCHOLARSHIPS AND AWARDS

2010	Thermo Fisher Scientific STEM Scholarship, Fudan University
2008, 2009	First Class Merit Scholarship, Fudan University
2008	Fosun Pharma Scholarship, Fudan University
2008	Fung Scholar, University of Hong Kong
2007	Chinese National Scholarship, Fudan University
2006	Tan Jiazhen (C. C. Tan) Life Sciences Scholarship, Fudan University